ISBN 978-85-89281-15-7

Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético

Marcos Deon Vilela de Resende

Embrapa

Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministério da Agricultura e do Abastecimento

Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético

Marcos Deon Vilela de Resende

Embrapa Florestas
Colombo, PR
2007

Exemplares desta publicação podem ser adquiridos na:

Embrapa Florestas
Estrada da Ribeira, Km 111, Guraituba,
83411 000 - Colombo, PR - Brasil
Caixa Postal: 319
Fone/Fax: (41) 3675 5600
Home page: www.cnpf.embrapa.br
E-mail: sac@cnpf.embrapa.br

Supervisão editorial: Luiz Roberto Graça
Revisão de texto: Mauro Marcelo Berté
Normalização bibliográfica: Lidia Woronkoff
Editoração eletrônica: Mauro Marcelo Berté
Capa: Mauro Marcelo Berté

1ª edição
1ª impressão (2007):

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Florestas

Resende, Marcos Deon Vilela de.

Matemática e estatística na análise de experimentos e no melhoramento genético / Marcos Deon Vilela de Resende. – Colombo : Embrapa Florestas, 2007.

362 p.

ISBN 978-85-89281-15-7

1. Matemática – Experimentação. 2. Estatística – Experimentação. I. Título.

CDD 001.422 (21. ed.)

Autor

Marcos Deon Vilela de Resende
Estatístico, Pós-Doutor,
Pesquisador da *Embrapa Florestas*,
marcos.deon@pq.cnpq.br

Apresentação

A presente publicação aborda ferramentas matemáticas e estatísticas com aplicações na análise de experimentos nas áreas agronômica, florestal e zootécnica. Contempla potenciais aplicações em espécies florestais, forrageiras, frutíferas, energéticas, estimulantes, palmáceas, olerícolas e espécies anuais produtoras de grãos e fibras. Abrange também a análise de dados no melhoramento para resistência a doenças e pragas, e os fundamentos dos métodos estatísticos descritos apresentam grande utilidade ao melhoramento animal.

Essa obra resulta da pesquisa científica conduzida pelo autor durante seus 20 anos de carreira profissional. Recomenda-se o seu uso juntamente com outras duas obras do autor: "*Genética Biométrica e Estatística no Melhoramento de Plantas Perenes (2002)" e "Selegen-Reml/Blup: Sistema Estatístico* e *Seleção Genética Computadorizada via Modelos Lineares Mistos (2007)".* Para aplicações práticas das metodologias descritas recomenda-se o uso do *software* Selegen-Reml/Blup, de autoria do mesmo autor.

Moacir José Sales Medrado
Chefe-Geral
Embrapa Florestas

Prefácio

Esta publicação apresenta algumas ferramentas matemáticas e estatísticas úteis na análise de experimentos de campo nas áreas agronômica, florestal e zootécnica, enfatizando métodos baseados em máxima verossimilhança. Contempla aplicações em espécies florestais, forrageiras, frutíferas, energéticas, olerícolas e também em espécies anuais produtoras de grãos e fibras.

A experimentação de campo apresenta certas peculiaridades, dentre as quais destacam-se: (i) desbalanceamento de dados devido a vários motivos como perdas de plantas e parcelas, desiguais quantidades de sementes e mudas disponíveis por tratamento, rede experimental com diferentes números de tratamentos e repetições por experimento e diferentes delineamentos experimentais, não avaliação de todas as combinações genótipo-ambiente, dentre outros; (ii) tendência ambiental ou variabilidade ambiental à pequena escala devido a fatores do solo tais como fertilidade, umidade, dentre outros; (iii) competição entre diferentes genótipos devido a variadas agressividades e sensibilidades dos diferentes materiais genéticos; (iv) heterogeneidade de variâncias entre e dentro de experimentos de uma rede experimental ou entre medidas repetidas tomadas sobre uma mesma unidade experimental.

Para contornar o problema exposto em (i), deve-se adotar o procedimento REML/BLUP (máxima verossimilhança residual/melhor predição linear não viciada), quando os efeitos de tratamentos forem considerados aleatórios, e o procedimento REML/GLS (máxima verossimilhança residual/quadrados mínimos generalizados), quando os efeitos de tratamentos forem considerados fixos. Esses procedimentos lidam naturalmente com o desbalanceamento conduzindo a estimações e predições mais precisas. O problema relatado em (ii) pode ser considerado por meio da análise estatística espacial de experimentos, principalmente via modelos de análise de séries temporais e métodos geoestatísticos aplicados via o procedimento REML e, também, por meio da adoção de adequados delineamentos e análises envolvendo blocos incompletos. A questão reportada em (iii) demanda o uso de modelos de competição intergenotípica visando contornar o problema da interferência de um tratamento sobre a resposta de outro. Em geral, os modelos de competição/interferência são aplicados em conjunto com a análise espacial e com o procedimento REML. A questão (iv) relativa à heterogeneidade de variâncias pode ser contornada pelo próprio REML, o qual permite especificar variâncias heterogêneas para os vários efeitos do modelo sob análise.

A análise de variância (ANOVA) e de regressão foram durante muito tempo os principais suportes da análise e modelagem estatística. Entretanto, essas técnicas têm como suposições básicas a independência dos erros e a independência entre os efeitos de tratamentos. O método REML permite relaxar esta suposição de independência permitindo maior flexibilidade na modelagem e também permitindo lidar com a situação de tratamentos correlacionados, o que é comum no melhoramento genético. Assim, o procedimento REML é o cerne deste documento. Tal procedimento foi criado pelos pesquisadores ingleses Robin Thompson e Desmond Patterson em 1971 e hoje constitui-se no procedimento padrão para a análise estatística em uma grande gama de aplicações.

Em experimentos agropecuários e florestais, o REML tem substituído com vantagens o método ANOVA criado pelo cientista inglês Ronald Fisher em 1925. Na verdade, o REML é uma generalização da ANOVA para situações mais complexas e pode também ser derivado e implementado sob o enfoque bayesiano. Para situações simples, os dois procedimentos são equivalentes. Mas para situações mais complexas e realísticas, a ANOVA é um procedimento apenas aproximado. Os algoritmos para obtenção de estimativas REML são essencialmente matemáticos. Após a criação do método, várias pesquisas foram conduzidas por Robin Thompson e outros autores visando à obtenção de algoritmos e programas computacionais mais eficientes.

Este livro aborda procedimentos ótimos de estimação e predição e também o software Selegen-Reml/Blup (descrito no livro *SELEGEN-REML/BLUP Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos*, o qual acompanha esse) como uma ferramenta simples e direta para uso rotineiro no dia a dia dos programas práticos de melhoramento de plantas anuais e perenes. No Brasil existem excelentes livros sobre Estatística e Experimentação Agronômica, todos eles fundamentados na ANOVA, ou seja, método de quadrados mínimos. Desta forma, acredita-se que esta publicação, fundamentada em REML e modelos mistos, possa, em parte, contribuir para complementar as obras existentes.

Marcos Deon Vilela de Resende

Curitiba, PR, 2007

Dedicatória

Às minhas filhas Annelise e Anna Carolina, por serem o melhor da minha vida. Muitos beijos para vocês.

Ao Professor Dr. Robin Thompson pela criação, aperfeiçoamento e implementação prática do método REML e pelos ensinamentos durante o meu Pós-Doutorado no Departamento de Matemática e Estatística do Rothamsted Research Institute, Inglaterra.

Sumário

CAPÍTULO 1
DELINEAMENTO DE EXPERIMENTOS DE CAMPO

1 ORIGEM E EVOLUÇÃO DOS MÉTODOS DE EXPERIMENTAÇÃO E ANÁLISE ESTATÍSTICA

Os fundamentos da moderna estatística teórica e experimental surgiram a partir do início do século 20, principalmente na Inglaterra. Grande parte dos desenvolvimentos realizados é devido a Ronald Fisher, que trabalhou na Rothamsted Experimental Station (hoje *Rothamsted Research Institute*) localizado ao norte de Londres. A Rothamsted Experimental Station foi fundada em 1843, como o pioneiro instituto de pesquisa agropecuária em termos mundiais. Por volta de 1920, o diretor do referido instituto decidiu contratar um matemático para criar um departamento de estatística no referido instituto, com o objetivo de analisar a grande massa de dados acumulados dos chamados "experimentos clássicos" instalados desde 1843. Fisher se candidatou ao cargo, foi contratado e permaneceu como chefe deste departamento por longo período. Durante este período, Fisher criou várias técnicas e conceitos que se tornaram centrais na ciência da Estatística. Em 1922 criou o método da máxima verossimilhança (Fisher, 1922). Em 1925 desenvolveu a análise de variância

(ANOVA), com implicações na estimação de componentes de variância e delineamentos experimentais (Fisher, 1925). Fisher (1926) enfatizou o papel crucial da repetição, da casualização e do controle local na eficiência dos experimentos. Criou então o delineamento em blocos completos casualizados. Na década de 1930, também em Rothamsted, Yates (1936; 1940) criou os delineamentos em blocos incompletos ou látice, os quais permitem um melhor controle da heterogeneidade experimental.

A partir dos fundamentos da análise matemática e estatística de experimentos de campo estabelecidos por Fisher e Yates, vários importantes livros tratando desse assunto foram publicados em várias partes do mundo, contribuindo para uma ampla divulgação desse tema. Considerando as línguas mais acessíveis aos brasileiros, ou seja, o inglês e aquelas de origem latina, citam-se as importantes obras pioneiras clássicas: (i) em inglês (Fisher, 1925; 1948; Kempthorne, 1952; Panse e Sukhatme, 1954; Federer, 1955; Steel e Torrie, 1960; Snedecor, 1967; Cochran e Cox, 1957); (ii) em português (Pimentel Gomes, 1966); (iii) espanhol (Garza, 1972); (iv) francês (Vessereau, 1960); (v) italiano (Vianelli, 1954); (vi) romeno (Ceapoiu, 1968; Giurgiu, 1966). As técnicas mais importantes até a época são bem descritas por Ceapoiu (1968), que apresenta também algumas contribuições russas escritas em romeno por Konstantinov e Plotnikov (1960), para atender a Moldávia, país ex-integrante da União Soviética e que tem como língua oficial o romeno. As técnicas de análise de experimentos foram estendidas também para outras áreas tecnológicas, por exemplo, para a química e indústria química (Bennet e Franklin, 1963).

Para uma análise mais eficiente de experimentos delineados em blocos incompletos, o procedimento máxima verossimilhança residual ou restrita (REML) foi criado pelos pesquisadores ingleses Robin Thompson e Desmond Patterson em 1971 e hoje constitui-se no procedimento padrão para a análise estatística em uma grande gama de aplicações. Em experimentos agronômicos e florestais, o REML tem substituído com vantagens o método ANOVA. O REML propicia grande flexibilidade na modelagem e permitiu o desenvolvimento e utilização da análise espacial (Cullis e Gleeson, 1991) e da simultânea análise espacial e da competição intergenotípica nos experimentos de campo (Stringer e Cullis, 2002; Resende e Thompson (2003); Resende et al., 2005).

Especificamente no caso do melhoramento de plantas, grande evolução nos métodos de análise ocorreu a partir da apresentação formal e completa do procedimento melhor predição linear não viciada (BLUP), conforme Henderson (1973; 1975) e Thompson (1976; 1979) para inferência sobre os efeitos genéticos de tratamentos. Tal procedimento é ótimo e, portanto, o mais preciso nas várias situações experimentais. Dessa forma passou a ser usado rotineiramente no melhoramento de plantas (HIll e Rosenberger, 1985; Resende et al., 1993; Resende et al., 1996; Bueno Filho, 1997; Duarte, 2000; Resende, 1999; 2000 e 2002). A análise REML/BLUP permitiu também maior eficiência na análise de medidas repetidas (Meyer e Hill, 1997; Dias e Resende, 2001; Resende e Thompson, 2003; Schaeffer, 2004) e na análise de múltiplos experimentos contemplando a estabilidade e adaptabilidade dos materiais genéticos (Phiepo, 1998; Smith et al., 2001; Resende e Thompson, 2003; 2004). Nesse último caso, os modelos fator analíticos mistos (FAMM) mostram-se mais adequados do que a técnica modelo de efeitos principais aditivos e de interações multiplicativos (AMMI).

2 PRINCÍPIOS BÁSICOS DA EXPERIMENTAÇÃO

Nesse capítulo são apresentados aspectos fundamentais na obtenção de dados experimentais fidedignos. Tais aspectos envolvem a escolha de um delineamento experimental adequado e o planejamento correto do tamanho de parcela e número de repetições a ser usado na experimentação.

Um delineamento experimental adequado deve obedecer aos princípios fundamentais da experimentação: repetição, casualização e controle local (Fisher, 1926). A importância do número de repetições é capital, significando que, com baixo número de repetições, até a casualização é prejudicada ou comprometida. Compromete-se também o número de graus de liberdade do resíduo. Menos que 8 graus de liberdade para o resíduo são insuficientes para se obter uma estimativa fidedigna do erro experimental ou variação residual (Mead, 1997). Com uma estimativa inadequada da variância residual, o teste F e todos os demais testes estatísticos não terão validade alguma. Nesse caso, selecionar ao acaso ou por sorteio os melhores tratamentos terá o mesmo efeito que selecionar por esses testes. O número de tratamentos em avaliação também influencia o número de graus de

liberdade do resíduo. Os números de tratamentos e de repetições e suas influências nos graus de liberdade do resíduo e na probabilidade de detecção de diferenças significativas entre tratamentos são fatores importantes na experimentação e têm sido muitas vezes esquecidos por ocasião do planejamento de experimentos.

A repetição refere-se ao número de vezes que o tratamento aparece no experimento. Tem por finalidade permitir a estimação do erro experimental, aumentar o poder dos testes estatisticos como o F e dos demais testes de médias e aumentar a precisão das estimativas das médias dos tratamentos. Neste último caso, quanto maior o número de repetições, menor é a variância da média dos tratamentos (Dias e Resende, 2001).

Como controle local, deve ser enfatizada a homogeneidade dentro de estratos ou blocos, sendo, em princípio, recomendados os delineamentos em blocos completos casualizados e em blocos incompletos (látice). A casualização e a repetição é que propiciam uma comparação não viciada dos tratamentos, ao passo que o controle local e a repetição permitem reduzir o erro experimental médio. Um erro experimental menor permite inferir como significativa uma diferença real pequena entre médias de tratamentos ou entre valores genéticos.

A casualização consiste em se dispor os tratamentos ao acaso no experimento de modo que todas as parcelas tenham a mesma chance de receber um determinado tratamento. É, portanto, recomendada para se evitar fatores sistemáticos que venham a beneficiar alguns tratamentos em detrimento de outros. Seu grande benefício é validar e dar confiabilidade às estimativas do erro experimental e das médias de tratamentos. O controle local destina-se a controlar a heterogeneidade ambiental e implica em restringir a casualização. Em termos da avaliação genética e da estimação de componentes de variância, a casualização é essencial como forma de evitar a correlação entre efeitos genéticos e ambientais, fato que afetaria todo o modelo básico de estimação e predição, o qual assume independência entre os referidos efeitos. Blocos nos delineamentos em blocos casualizados e linhas e colunas nos quadrados latinos, por exemplo, são estratégias de controle local que possibilitam agrupar parcelas homogêneas e casualizar os tratamentos dentro deles. O uso da análise de covariância é também um tipo de controle local e é muitas vezes denominado também de controle estatístico (Dias e Resende, 2001).

O delineamento em quadrado latino propicia melhor controle local, visto que permite controlar a heterogeneidade ambiental em duas direções, no sentido das linhas e das colunas. Entretanto, tal delineamento não tem sido recomendado para os trabalhos de melhoramento (Ramalho et al., 2000) ou na experimentação em geral (Pimentel Gomes, 1987), devido à restrição do número de repetições ter que ser igual ao número de tratamentos ou progênies. Dessa forma, não há relatos de sua utilização no melhoramento. No entanto, com o advento da utilização de parcelas de uma planta no melhoramento de plantas perenes, tal delineamento passa a ter grande potencial de utilização. Como se utiliza em torno de 60 plantas por progênie (60 repetições de uma planta), quadrados latinos de 60 x 60 com 60 progênies poderiam ser perfeitamente utilizados, em associação com o procedimento BLUP (melhor predição linear não viciada). No caso, os dados seriam corrigidos para dois gradientes ambientais (linhas e colunas), pelo método BLUP ou do índice multiefeitos derivado por Resende e Higa (1994) e extendido para o delineamento em linha e coluna, conforme apresentado no Capítulo 5. Segundo Panse e Sukhatme (1963), quando existem tendências simultâneas de variações em fertilidade em duas direções em ângulos retos (que equivale a uma tendência diagonal em fertilidade), é provável que o quadrado latino seja mais eficiente que o delineamento em blocos. O delineamento em quadrado latino é também recomendado quando não se conhece a priori os gradientes de fertilidade. Uma generalização do quadrado latino é o delineamento em linha e coluna, o qual, atualmente, pode ser delineado com qualquer número de repetições e ser analisado via metodologia de modelos mistos (REML/BLUP).

Os delineamentos de blocos incompletos (látice, por exemplo) são especialmente indicados na situação de grande número de tratamentos e alta variabilidade ambiental na área experimental. A eficiência relativa entre os delineamentos experimentais depende, sobretudo, do nível de variação ambiental espacial na área experimental. Empregando um modelo geoestatístico espacial, o qual permite a especificação de vários níveis de variação ambiental, Fu et al. (1998) concluíram pela superioridade dos delineamentos de blocos incompletos (látice e alfa) em um grande número de situações, em termos de eficiência estatística para a estimação de médias de tratamentos.

Outro delineamento bastante usado é o de blocos aumentados, proposto por Federer (1958), o qual não apresenta repetições dos tratamentos principais. Esse delineamento é, por construção, desbalanceado e não ortogonal. Assim, deve ser analisado via metodologia de modelos mistos.

Tal delineamento é útil nas fases iniciais do melhoramento, quando não se dispõe de propágulos suficientes dos materiais genéticos para fins de experimentação.

Delineamentos experimentais ótimos podem ser construídos usando a matriz de informação dos parâmetros e sua inversa. Para o modelo y = Xb + Za + e, associado ao vetor de observações (y), vetor de efeitos de blocos (b), vetor de efeitos de tratamentos (a) e efeitos de erros não correlacionados (e), a matriz de informação equivale a $M = (1/\sigma^2)\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma^2 G^{-1} \end{bmatrix}$, ou seja, é a própria matriz dos coeficientes das equações de modelo misto. X e Z são as matrizes de incidência para b e a, respectivamente, e G é a matriz de covariância associada ao vetor a. A matriz inversa de M dada por $M^{-1} = \sigma^2 \begin{bmatrix} C^{XX} & C^{XZ} \\ C^{ZX} & C^{ZZ} \end{bmatrix}$ é uma matriz de covariância de amostragem dos efeitos estimados. O critério de optimalidade deve minimizar (ou maximizar) alguma função da matriz C^{zz} associada aos efeitos (aleatórios no caso) de tratamentos (a). Chaloner e Verdinelli (1995) consideram com detalhe este tema.

Nesse livro são abordados procedimentos de análise de experimentos estabelecidos nos delineamentos inteiramente ao acaso, blocos completos, blocos incompletos, blocos aumentados e linha e coluna.

3 TAMANHO DE PARCELA

Vários estudos realizados confirmam a maior eficiência dos delineamentos com uma planta ou observação por parcela em relação àqueles com várias plantas por parcela. Esta superioridade advém de:

a) maior precisão na comparação entre tratamentos devido ao maior número de repetições para uma área experimental de tamanho fixo;

b) maior acurácia seletiva devido ao aumento do número de repetições, em uma área experimental de tamanho fixo;

c) maior herdabilidade individual no bloco devido à obtenção de blocos mais homogêneos;

d) atenuação dos efeitos de competição devido à ocorrência de maior número de diferentes vizinhos;

e) menor superestimativa (devido à interação genótipos x ambientes) da herdabilidade e do ganho genético em um local, pois são utilizados maiores números de repetições (que podem representar diferentes ambientes);

f) menor tamanho e maior homogeneidade do bloco, reduzindo a necessidade de análise espacial de experimentos, pois o controle local é mais efetivo;

Como comprovação prática, as parcelas com uma planta ou observação têm sido empregadas no melhoramento florestal em várias partes do mundo. No Brasil, são usadas, por exemplo, em algumas empresas de celulose, em testes clonais e de progênies, onde constataram-se: aumento da herdabilidade, aumento da acurácia seletiva, ausência de efeitos de competição. A tecnologia de seleção atualmente disponível (REML/BLUP) baseia-se no valor genético predito, o qual é obtido após rigorosa correção para todos os efeitos ambientais. Assim, ocorre ajuste para a variação físico-química do solo, por meio do ajuste para os efeitos de blocos, parcelas e locais. Com essa tecnologia também não há problema de análise, quando se perde a representação de alguns tratamentos em alguns blocos, devido à mortalidade das plantas.

Empregando a metodologia do coeficiente de correlação intraclasse (relação variância entre parcelas/(variância entre parcelas + variância dentro de parcelas)), Pimentel Gomes (1984) concluiu pelo uso de uma planta por parcela. Considerando a precisão experimental e a probabilidade de detecção de diferenças significativas entre médias de tratamentos, Cotterill e James (1984), Loo-Dinks e Tauer (1987), e Haapanen (1992) também concluíram pelo uso de uma planta por parcela. Maiores detalhes referentes ao tamanho ideal de parcela são apresentados por Ramalho et al. (2000) e Resende (2002).

O conceito de parcelas de uma planta aplica-se igualmente a cana-de-açúcar e forrageiras. Em cana-de-açúcar e capim elefante, isto deve ser interpretado como uma observação por parcela,

ou seja, um sulco por parcela, nas situações em que se colhe todo o sulco como mistura para se proceder à avaliação. No caso, define-se a herdabilidade individual como herdabilidade associada a um sulco e o número de repetições é determinado em função da magnitude desta herdabilidade ao nível de um sulco. No melhoramento de forrageiras, o uso de bordaduras externas ao experimento é fundamental. Bordaduras internas nas próprias parcelas podem também ser usadas eficientemente em alguns casos. O uso de maiores espaçamentos entre linhas apresenta a vantagem de evitar competição diferenciada entre tratamentos, mas conduz a superestimativas das produções, pois não há competição lateral. No entanto, para o objetivo de ordenamento de genótipos para seleção, recomenda-se o uso de experimentação com plantas mais espaçadas entre si, de forma a evitar a sobreposição das mesmas.

4 TAMANHO AMOSTRAL E NÚMERO DE REPETIÇÕES

4.1 Tamanho Amostral para Seleção Genética e Estimação de Componentes de Variância

Tamanhos amostrais (em termos do número de repetições) para estimação e predição em várias situações práticas no melhoramento genético são detalhados por Resende (2002) e Resende e Barbosa (2005). Para determinação dos tamanhos amostrais, tais autores adotaram o critério de maximização da acurácia seletiva com o aumento do número de repetições.

4.2 Tamanho Amostral para Estimação da Média Populacional

A abordagem apresentada a seguir foi desenvolvida por Stein (1945) e concentra-se mais em estimação do que em testes de hipóteses.

A abordagem refere-se ao uso do intervalo de confiança para a média amostral ($\bar{y}$), dada por: $I.C. = \bar{y} \pm z\ s(\bar{y}) = \bar{y} \pm 1{,}96\ s(\bar{y})$, em que $s(\bar{y}) = [\hat{\sigma}^2 / N]^{1/2}$ é o erro padrão da média. Neste caso, pode-se estabelecer um erro tolerável (δ) na estimativa da média, dado por $\delta = 1{,}96\ s(\bar{y}) = 1{,}96\ [\hat{\sigma}^2 / N]^{1/2}$. A partir desta expressão, obtém-se $N = \frac{(1{,}96)^2 \hat{\sigma}^2}{\delta^2}$, como o tamanho adequado da amostra, para um erro

tolerável δ, escolhido a priori. Com σ^2 desconhecido, utiliza-se a estimativa $\hat{\sigma}^2$ e o valor (1,96) t da distribuição de Student em vez do z da distribuição normal.

Assim, para determinação do tamanho amostral N, necessita-se estabelecer um erro tolerável (δ) na estimativa da média e ter uma estimativa $\hat{\sigma}^2$ da variabilidade fenotípica na população. O erro δ pode ser especificado em porcentagem da média, por exemplo, 10 %. Neste caso δ é dado por $0{,}10\ \bar{y}\ e\ N = \frac{(1{,}96)^2\ \hat{\sigma}^2}{(0{,}10\ \bar{y})^2} = \frac{(1{,}96)^2\ \hat{\sigma}^2}{(0{,}10)^2\ \bar{y}^2} = \frac{(1{,}96)^2}{(0{,}10)^2}\ (CV)^2$, em que $CV = \frac{\hat{\sigma}}{\bar{y}}$ é o coeficiente de variação do caráter na população.

Assim, por esta abordagem, torna-se necessária apenas uma estimativa ou conhecimento prévio do coeficiente de variação fenotípico individual na população e quanto maior este, maior será o tamanho necessário da amostra. Considerando aceitável um erro de 10 %, obtém-se os seguintes tamanhos adequados da amostra em função de diferentes coeficientes de variação na população (Tabela 1).

Tabela 1. Amostragem adequada por população, quantificada pelo número total de indivíduos (*n*), em função de alguns coeficientes de variação fenotípico individual na população, visando a um erro aceitável de 10 % na estimativa da média populacional.

CV	N	CV	N
0,10	4	0,50	96
0,20	15	0,60	138
0,30	35	0,70	188
0,40	61	0,80	246

Para variáveis binomiais e usando-se a aproximação à distribuição normal, a mesma expressão para N pode ser utilizada, bastando substituir $\hat{\sigma}^2$ por pq e $\bar{y}$ por p, em que p e q referem-se às proporções observadas das duas classes fenotípicas. Neste caso, torna-se necessária apenas uma estimativa prévia de p, a proporção média observada, visto que q = 1-p.

4.3 Tamanho Amostral para Detecção de Diferenças Significativas entre Tratamentos

Informações referentes ao tamanho amostral adequado são essenciais no planejamento e execução da experimentação científica. A presente abordagem objetiva fornecer os tamanhos amostrais adequados em função do coeficiente de determinação (ou herdabilidade) dos efeitos de tratamentos.

Os livros texto de estatística fornecem a seguinte expressão para cálculo do tamanho amostral (N) adequado (Snedecor e Cochran, 1967; Steel e Torrie, 1980):

$N = \frac{(z_\alpha + z_\beta)^2 \sigma_D^2}{\delta^2}$ em que:

z_∂ e z_β: valores da função distribuição acumulada da distribuição normal padrão, associados às probabilidades de erro tipo I (α) e erro tipo II (β), em teste de hipótese unilateral;

σ_D^2: variância da diferença entre duas médias de tratamentos;

δ: tamanho da diferença verdadeira entre duas médias, que se deseja discriminar como significativa.

Para testes bilaterais a expressão de N passa a ser:

$$N = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_D^2}{\delta^2}$$

A quantidade $(1-\beta)$ é equivalente a P, a probabilidade de que o experimento exiba uma diferença estatisticamente significativa entre médias de tratamentos. Probabilidades de 0,80 e 0,90 são comuns e adequadas na prática. Probabilidades maiores tais quais 0,95 ou 0,99, podem ser consideradas, mas os tamanhos amostrais necessários para atingir tais especificações são anti-econômicos (Snedecor e Cochran, 1967).

Valores de $(z_\alpha + z_\beta)^2$ e $(z_{\alpha/2} + z_\beta)^2$ foram apresentados por Snedecor e Cochran (1967) e são apresentados na seqüência:

P	Testes Bilaterais $(z_{\alpha/2} + z_\beta)^2$			Testes Unilaterais $(z_\alpha + z_\beta)^2$		
	Nível de Significância α			Nível de Significância α		
	0,01	**0,05**	**0,1**	**0,01**	**0,05**	**0,1**
0,80	11,7	7,9	6,2	10,0	6,2	4,5
0,90	14,9	10,5	8,6	13,0	8,6	6,6
0,95	17,8	13,0	10,8	15,8	10,8	8,6

Verifica-se que os multiplicadores de $\frac{\sigma_D^2}{\delta^2}$ para os testes unilaterais ao nível de 5 % de significância equivalem aos multiplicadores para os testes bilaterais ao nível de 10 % de significância. Como ilustração, o valor de $(z_\alpha + z_\beta)^2$ = 8,6 para o teste unilateral ao nível de 5 % de significância e com probabilidade de 90 % de obtenção de diferenças significativas entre tratamentos é dado por $[(z_\alpha = z_{0,05} = 1{,}645) + (z_\beta = z_{1-p} = z_{0,10} = 1{,}282)]^2$ = $(1{,}645+1{,}282)^2 = 8{,}57 \approx 8{,}6$.

A variância σ_D^2 da diferença entre médias para amostras independentes é dada por $\sigma_D^2 = 2\sigma^2$, em que σ^2 é a variância residual da população ou variância do erro. Em testes de progênies σ^2 equivale a $(1-\rho_a h_a^2)\sigma_y^2$, para progênies de meios irmãos, e $1-[\rho_a h_a^2 + \rho_d (h_g^2 - h_a^2)]\sigma_y^2$, para progênies de irmãos germanos (Resende, 2002), em que:

$\rho_a \, e \, \rho_g$: correlação genética aditiva e de dominância entre indivíduos das progênies, respectivamente. O coeficiente ρ_a equivale a 0,25 para meios irmãos e 0,5 para irmãos germanos e ρ_d equivale a 0,25 para irmãos germanos;

$h_a^2 \, e \, h_g^2$: herdabilidade individual no sentido restrito e amplo, respectivamente;

σ_y^2 : variância fenotípica da população.

Para progênies de meios irmãos, $\sigma_D^2 = 2\sigma^2 = 2(1-0{,}25h_a^2)\sigma_y^2$ e a expressão do tamanho amostral torna-se:

$$N = \frac{(z_\alpha + z_\beta)^2 2(1-0{,}25h_a^2)\sigma_y^2}{\delta^2} = \frac{(z_\alpha + z_\beta)^2 (2-0{,}5h_a^2)\sigma_y^2}{\delta^2}$$

Assim, para determinação do tamanho amostral, é necessário:

a) Uma estimativa de σ^2, no caso representada por h_a^2 e σ_y^2;

b) Escolha de uma probabilidade P de obtenção de diferenças significativas entre médias de tratamentos;

c) Escolha do nível de significância do teste unilateral ou bilateral;

d) Tamanho da diferença verdadeira δ que se deseja detectar como significativa.

A diferença δ pode ser definida em termos de desvios padrões fenotípicos e, então, ser expressa como $\frac{\delta}{\sigma_y}$, a mínima diferença em desvios padrões fenotípicos que se deseja detectar como significativa.

Trabalhando-se a equação para N, obtêm-se:

$$\frac{\delta^2}{\sigma_y^2} = \frac{(z_\alpha + z_\beta)^2 (2-0{,}5h_a^2)}{N}$$

A mínima diferença $\frac{\delta}{\sigma_y}$ decresce assintoticamente com o aumento de N, de forma que pode-se determinar N que praticamente minimiza $\frac{\delta}{\sigma_y}$.

Tomando-se como adequado P = 0,90 e α = 5 % para o teste unilateral (ou α=10 % para o teste bilateral), tem-se $(z_\alpha + z_\beta)^2$ = 8,6 e a função a ser minimizada é dada por $\frac{\delta}{\sigma_y} = \left[\frac{8{,}6(2-0{,}5h_a^2)}{N}\right]^{1/2} = \left[\frac{(17{,}2-4{,}3h_a^2}{N}\right]^{1/2}$.

Esta expressão deve ser trabalhada para diferentes valores de h_a^2 e para valores crescentes de N.

Para experimentos envolvendo tratamentos genéricos, tais como variedades, linhagens, híbridos, clones, procedências, espécies, a correlação intraclasse equivale a h_g^2 e $\sigma_D^2 = 2(1-h_g^2)\sigma_y^2$. Neste caso, a expressão para N passa a ser dada por:

$$N = \frac{(z_\alpha + z_\beta)^2 2(1-h_g^2)\sigma_y^2}{\delta^2} = \frac{(z_\alpha + z_\beta)^2 (2-2h_g^2)\sigma_y^2}{\delta^2}$$

e a função a ser minimizada é dada por:

$$\frac{\delta}{\sigma_y} = \left[\frac{8{,}6(2-2h_g^2)}{N}\right]^{1/2} = \left[\frac{17{,}2(1-h_g^2)}{N}\right]^{1/2}$$

A quantidade h_g^2 representa genericamente a herdabilidade individual dos efeitos de tratamentos, dada por $h_g^2 = \frac{\sigma_t^2}{\sigma_y^2}$, sendo que σ_t^2 é variância entre tratamentos. De forma mais genérica, para contemplar também os tratamentos de efeitos fixos e outros tratamentos não genéticos, h_g^2 representa um coeficiente de determinação individual (R^2) dos efeitos de tratamento.

Esta abordagem pode ser aplicada também quando tratamentos são analisados ao nível de médias de parcela. Nesta situação h_g^2 refere-se à herdabilidade (ou coeficiente de determinação) de tratamentos ao nível de parcela e o valor N refere-se ao número de repetições a ser adotado. No caso de testes de progênies, N refere-se ao número de repetições para o caso de uma planta por parcela ou avaliação ao nível de média de parcela. Na Tabela 2 são apresentados os tamanhos amostrais necessários em várias situações.

Tabela 2. Tamanho amostral (n), em função do coeficiente de determinação individual (R^2) dos efeitos de tratamento, para detecção de diferenças significativas entre tratamentos em número de desvios padrões fenotípicos (δ / σ_y).

h^2 ou R^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.05	5	1.84	1.81
0.05	10	1.30	1.28
0.05	15	**1.06**	**1.04**
0.05	20	0.92	0.90
0.05	30	0.75	0.74
0.05	40	0.65	0.64
0.05	50	0.58	0.57
0.05	60	0.53	**0.52**
0.05	70	**0.49**	0.48
0.05	80	0.46	0.45
0.05	90	0.43	0.43
0.05	100	0.41	0.40
0.1	5	1.83	1.76
0.1	10	1.29	1.24
0.1	15	**1.06**	**1.02**
0.1	20	0.92	0.88
0.1	30	0.75	0.72
0.1	40	0.65	0.62
0.1	50	0.58	0.56
0.1	60	0.53	**0.51**
0.1	70	**0.49**	0.47
0.1	80	0.46	0.44
0.1	90	0.43	0.41
0.1	100	0.41	0.39
0.15	5	1.82	1.71
0.15	10	1.29	1.21
0.15	15	**1.05**	**0.99**
0.15	20	0.91	0.85
0.15	30	0.74	0.70
0.15	40	0.64	0.60
0.15	50	0.58	0.54

continuação Tabela 2

h^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.15	60	0.53	**0.49**
0.15	70	**0.49**	0.46
0.15	80	0.45	0.43
0.15	90	0.43	0.40
0.15	100	0.41	0.38
0.2	5	1.81	1.66
0.2	10	1.28	1.17
0.2	15	**1.04**	**0.96**
0.2	20	0.90	0.83
0.2	30	0.74	0.68
0.2	40	0.64	0.59
0.2	50	0.57	**0.52**
0.2	60	**0.52**	0.48
0.2	70	0.48	0.44
0.2	80	0.45	0.41
0.2	90	0.43	0.39
0.2	100	0.40	0.37
0.25	5	1.80	1.61
0.25	10	1.27	1.14
0.25	15	**1.04**	**0.93**
0.25	20	0.90	0.80
0.25	30	0.73	0.66
0.25	40	0.63	0.57
0.25	50	0.57	**0.51**
0.25	60	**0.52**	0.46
0.25	70	0.48	0.43
0.25	80	0.45	0.40
0.25	90	0.42	0.38
0.25	100	0.40	0.36
0.3	5	1.78	1.55
0.3	10	1.26	**1.10**
0.3	15	**1.03**	**0.90**
0.3	20	0.89	0.78

continuação Tabela 2

h^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.3	30	0.73	0.63
0.3	40	0.63	0.55
0.3	50	0.56	**0.49**
0.3	60	**0.51**	0.45
0.3	70	0.48	0.41
0.3	80	0.45	0.39
0.3	90	0.42	0.37
0.3	100	0.40	0.35
0.35	5	1.77	1.50
0.35	10	1.25	**1.06**
0.35	15	**1.02**	0.86
0.35	20	0.89	0.75
0.35	30	0.72	0.61
0.35	40	0.63	**0.53**
0.35	50	0.56	0.47
0.35	60	**0.51**	0.43
0.35	70	0.47	0.40
0.35	80	0.44	0.37
0.35	90	0.42	0.35
0.35	100	0.40	0.33
0.4	5	1.76	1.44
0.4	10	1.24	**1.02**
0.4	15	**1.02**	0.83
0.4	20	0.88	0.72
0.4	30	0.72	0.59
0.4	40	0.62	**0.51**
0.4	50	0.56	0.45
0.4	60	**0.51**	0.41
0.4	70	0.47	0.38
0.4	80	0.44	0.36
0.4	90	0.41	0.34
0.4	100	0.39	0.32

continuação Tabela 2

h^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.45	5	1.75	1.38
0.45	10	1.24	**0.97**
0.45	15	**1.01**	0.79
0.45	20	0.87	0.69
0.45	30	0.71	0.56
0.45	40	0.62	**0.49**
0.45	50	0.55	0.43
0.45	60	**0.50**	0.40
0.45	70	0.47	0.37
0.45	80	0.44	0.34
0.45	90	0.41	0.32
0.45	100	0.39	0.31
0.5	5	1.73	1.31
0.5	10	1.23	**0.93**
0.5	15	**1.00**	0.76
0.5	20	0.87	0.66
0.5	30	0.71	**0.54**
0.5	40	0.61	0.46
0.5	50	0.55	0.41
0.5	60	**0.50**	0.38
0.5	70	0.46	0.35
0.5	80	0.43	0.33
0.5	90	0.41	0.31
0.5	100	0.39	0.29
0.55	5	1.72	**1.24**
0.55	10	1.22	**0.88**
0.55	15	**0.99**	0.72
0.55	20	0.86	0.62
0.55	30	0.70	**0.51**
0.55	40	0.61	0.44
0.55	50	0.54	0.39
0.55	60	**0.50**	0.36

continuação Tabela 2

h^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.55	70	0.46	0.33
0.55	80	0.43	0.31
0.55	90	0.41	0.29
0.55	100	0.39	0.28
0.6	5	1.71	**1.17**
0.6	10	1.21	**0.83**
0.6	15	**0.99**	0.68
0.6	20	0.85	0.59
0.6	30	0.70	**0.48**
0.6	40	0.60	0.41
0.6	50	0.54	0.37
0.6	60	**0.49**	0.34
0.6	70	0.46	0.31
0.6	80	0.43	0.29
0.6	90	0.40	0.28
0.6	100	0.38	0.26
0.7	5	1.68	**1.02**
0.7	10	1.19	0.72
0.7	15	**0.97**	0.59
0.7	20	0.84	**0.51**
0.7	30	0.69	0.41
0.7	40	0.60	0.36
0.7	50	0.53	0.32
0.7	60	**0.49**	0.29
0.7	70	0.45	0.27
0.7	80	0.42	0.25
0.7	90	0.40	0.24
0.7	100	0.38	0.23
0.8	5	1.66	0.83
0.8	10	1.17	0.59
0.8	15	**0.96**	**0.48**
0.8	20	0.83	0.41

conclusão Tabela 2

h^2	N	Meios-irmãos δ / σ_y	Trat. Genéricos δ / σ_y
0.8	30	0.68	0.34
0.8	40	0.59	0.29
0.8	50	**0.52**	0.26
0.8	60	0.48	0.24
0.8	70	0.44	0.22
0.8	80	0.41	0.21
0.8	90	0.39	0.20
0.8	100	0.37	0.19
0.9	5	1.63	**0.59**
0.9	10	1.15	0.41
0.9	15	**0.94**	0.34
0.9	20	0.82	0.29
0.9	30	0.67	0.24
0.9	40	0.58	0.21
0.9	50	**0.52**	0.19
0.9	60	0.47	0.17
0.9	70	0.44	0.16
0.9	80	0.41	0.15
0.9	90	0.38	0.14
0.9	100	0.37	0.13

Na Tabela 2 estão assinalados os tamanhos amostrais necessários para a detecção de diferenças da ordem de aproximadamente um desvio padrão e meio desvio padrão entre médias de tratamentos. Verifica-se que para materiais mais variáveis geneticamente (como progênies de meios irmãos) os tamanhos amostrais requeridos são maiores. O tamanho amostral necessário diminui com o aumento da herdabilidade ou R^2 da variável, mas esta redução é pequena.

Considerando o objetivo de se detectar diferenças de magnitude equivalente a um desvio padrão da característica, tem-se: (i) para progênies de meios irmãos, o tamanho amostral necessário varia em torno de 15 para as diferentes herdabilidades; (ii) para clones, populações e tratamentos genéricos, o tamanho amostral necessário varia de 15 para a herdabilidade de 5 % a 5 para a

herdabilidade de 70 %.

Se o objetivo for a detecção de diferenças menores (da ordem de 0,5 desvio padrão) entre médias de tratamentos, maiores tamanhos amostrais são requeridos. Estes tamanhos são: (i) para progênies de meios irmãos, varia de 70 a 50 para as herdabilidades ou R^2 de 5 % a 90 %, respectivamente; (ii) para clones, populações e tratamentos genéricos, varia de 60 a 5 para as herdabilidades ou R^2 de 5 % a 90 %, respectivamente.

É importante ressaltar que a abordagem apresentada é válida para o caso em que se deseja avaliar tratamentos assumidos como de efeitos fixos. Isto porque a derivação da expressão básica apresentada por Steel e Torrie (1980) e Snedecor e Cochran (1967) baseia-se em testes de hipótese, cuja teoria considera os efeitos de tratamentos como fixos. Entretanto, isto não impede que informações intrínsecas à variável analisada, tais quais a herdabilidade (obtidas em estudos considerando os efeitos de tratamentos como aleatórios) possam ser usadas no planejamento dos experimentos envolvendo tratamentos de efeitos fixos. Neste caso, h^2 desempenha o papel de R^2 nas expressões apresentadas.

5 PRINCIPAIS DELINEAMENTOS E ARRANJOS EXPERIMENTAIS: MODELOS, GRAUS DE LIBERDADE E ESPERANÇA DE QUADRADOS MÉDIOS

O termo delineamento experimental está diretamente relacionado à forma de controle local (também denominado restrição na casualização) e o termo delineamento de tratamentos refere-se à forma de organizar os tratamentos em estruturas ou arranjos de classificação cruzada (fatorial) ou hierárquica. Experimentos em parcelas subdivididas podem ser referidos como uma forma de arranjo (delineamento de tratamentos) e também como um delineamento experimental, nesse caso, como um tipo especial de blocos incompletos. Os experimentos em parcelas subdivididas são caracterizados por uma restrição adicional na casualização.

Nesse tópico são considerados os principais delineamentos experimentais (inteiramente ao acaso, blocos ao acaso, quadrado latino e látice) e arranjos experimentais de acordo com a estrutura cruzada ou hierárquica dos tratamentos (fatorial, parcelas subdivididas e hierárquica). Nesse último

caso, considera-se que os experimentos foram instalados no delineamento de blocos ao acaso. Detalhes práticos desses delineamentos são apresentados por Campos (1984), Gomes (1987), Banzato e Kronka (1989), Sampaio (1998), Ramalho et al. (2000), Storck et al. (2000), Dias e Resende (2001) e Barbin (2003), dentre outros.

Os esquemas de análise apresentados são importantes tanto no contexto da análise de variância tradicional (ANAVA) com dados balanceados usando o método de quadrados mínimos quanto no contexto da análise de deviance (ANADEV) para dados desbalanceados usando máxima verossimilhança residual (REML), conforme descrito nos Capítulos 3, 4 e 10. Tais esquemas informam sobre a estrutura dos fatores, sobre o número de graus de liberdade para ajuste dos vários efeitos e sobre a composição dos fatores e resíduos em termos de componentes de variância, o que auxilia na realização de testes de hipóteses adequados.

As esperanças matemáticas de quadrados médios propriamente ditas não têm importância e utilidade nos estimadores de componentes de variância via REML. No entanto, podem ajudar a entender a composição de algumas variâncias residuais associadas ao ajuste de alguns modelos. A obtenção de esperanças de quadrados médios é apresentada com detalhes por Bennet e Franklin (1963) e Hicks (1973). Os diferentes fatores podem ser considerados como de efeitos fixos ou aleatórios. As interações entre os fatores são também de efeitos fixos ou aleatórios e dependem das suposições dos efeitos dos fatores principais. Para fatores de classificação cruzada, as interações serão de efeitos fixos se todos os fatores principais forem de efeitos fixos e serão de efeitos aleatórios se pelo menos um dos fatores principais for de efeitos aleatórios.

5.1 Delineamento Inteiramente ao Acaso

Modelo Estatístico

$Y_{ij} = u + t_i + e_{ij}$, em que Y_{ij} é o valor medido no campo, u é a média geral, t_i é o efeito do tratamento i e e_{ij} é o efeito do erro aleatório ou resíduo.

a) Fator tratamentos de efeitos fixos.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Tratamentos	t-1	Q_t	$\sigma^2 + [b/(t-1)]\sum_i t_i^2$	Q_t/Q_r
Erro	(b-1)t	Q_r	σ^2	-

b) Fator tratamentos de efeitos aleatórios.

Fontes de Variação	**GL**	**QM**	**E(QM)**	**F sob H_o**
Tratamentos	t-1	Q_t	$\sigma^2 + b\sigma_t^2$	Q_t/Q_r
Erro	(b-1)t	Q_r	σ^2	-

Verifica-se que a consideração dos efeitos de tratamentos como fixos ou aleatórios não afeta a composição do teste F. As letras t e b denotam número de tratamentos e de repetições, respectivamente.

5.2 Delineamento em Blocos ao Acaso

Modelo Estatístico

$Y_{ij} = u + t_i + b_j + e_{ij}$, em que Y_{ij} é o valor medido no campo, u é a média geral, t_i é o efeito do tratamento i, b_j é o efeito do bloco j e e_{ij} é o efeito do erro aleatório ou resíduo.

a) Fator tratamentos de efeitos fixos e fator bloco de efeitos aleatórios.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Blocos	b-1	Q_b	$\sigma^2 + t\sigma_b^2$	Q_b/Q_r
Tratamentos	t-1	Q_t	$\sigma^2 + [b/(t-1)]\sum_i t_i^2$	Q_t/Q_r
Erro	(b-1)(t-1)	Q_r	σ^2	-

b) Fator tratamentos de efeitos aleatórios e fator bloco de efeitos aleatórios.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Blocos	b-1	Q_b	$\sigma^2 + t\sigma_b^2$	Q_b/Q_r
Tratamentos	t-1	Q_t	$\sigma^2 + b\sigma_t^2$	Q_t/Q_r
Erro	(b-1)(t-1)	Q_r	σ^2	-

Verifica-se que a consideração dos efeitos de tratamentos como fixos ou aleatórios não afeta a composição do teste F.

5.3 Delineamento em Quadrado Latino

Modelo Estatístico

$Y_{(i)jk} = u + t_{(i)} + l_j + c_k + e_{(i)jk}$, em que $Y_{(i)jk}$ é o valor medido no campo no tratamento i na parcela jk, u é a média geral, $t_{(i)}$ é o efeito do tratamento i, l_j é o efeito do linha j, c_k é o efeito da coluna k e $e_{(i)jk}$ é o efeito do erro aleatório ou resíduo associado à parcela jk dada pela combinação da linha j com a coluna k, parcela essa em que foi aplicado o tratamento i.

a) Fator tratamentos de efeitos fixos e fatores linha e coluna de efeitos aleatórios.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Linhas	l-1	Q_l	$\sigma^2 + b\sigma_l^2$	Q_l/Q_r
Colunas	c-1	Q_c	$\sigma^2 + b\sigma_c^2$	Q_c/Q_r
Tratamentos	t-1	Q_t	$\sigma^2 + [b/(t-1)]\sum_i t_{(i)}^2$	Q_t/Q_r
Erro	(b-1)(b-2)	Q_r	σ^2	-

No delineamento em quadrado latino, os números de tratamentos (t), de repetições (b), de linhas (l) e de colunas (c) são todos iguais, ou seja, t = b = l = c. Na verdade, linhas e colunas significam tratamentos e repetições. Assim, o número total de parcelas no experimento é dado pelo quadrado do número de repetições ou de tratamentos.

b) Fatores tratamentos, linhas e colunas de efeitos aleatórios.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Linhas	l-1	Q_l	$\sigma^2 + b\sigma_l^2$	Q_l/Q_r
Colunas	c-1	Q_c	$\sigma^2 + b\sigma_c^2$	Q_c/Q_r
Tratamentos	t-1	Q_t	$\sigma^2 + b\sigma_t^2$	Q_t/Q_r
Erro	(b-1)(b-2)	Q_r	σ^2	-

Verifica-se que a consideração dos efeitos de tratamentos como fixos ou aleatórios não afeta a composição do teste F. O delineamento em quadrado latino é o precursor do delineamento em linha e coluna, no qual não há a restrição do número de tratamentos ser igual ao número de repetições.

5.4 Delineamento em Látice Quadrado Balanceado

Modelo Estatístico

$Y_{ijk} = u + t_i + r_j + b_{jk} + e_{ijk}$, em que Y_{ijk} é o valor medido no campo, u é a média geral, t_i é o efeito do tratamento i, r_j é o efeito da repetição j, b_{jk} é o efeito do bloco k dentro da repetição j e e_{ijk} é o efeito do erro aleatório ou resíduo associado à cada parcela.

a) Fator tratamentos de efeitos fixos e fator bloco dentro de repetição de efeitos aleatórios.

Para a análise com recuperação da informação inter-blocos tem-se:

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	-	-
Blocos Ajustados dentro Repet.	(k-1)r	Q_b	$\sigma^2 + k\sigma_b^2$	-
Tratamentos	t-1	Q_t	$\sigma^2[1+\sigma_b^2/(\sigma^2+k\sigma_b^2)] + [r/(t-1)]\sum_i t_i^2$	Q_t/Q_{EE}
Erro Efetivo	(k-1)(t-1)	Q_{EE}	$\sigma^2[1+\sigma_b^2/(\sigma^2+k\sigma_b^2)]$	-
Erro Intrabloco	(k-1) (t-1)	Q_{EI}	σ^2	-

No delineamento em látice quadrado balanceado, o número de tratamento por bloco dentro de repetição é denotado por k e o número total de tratamentos é dado por $t = k^2$.

b) Fatores tratamentos e blocos dentro de repetição de efeitos aleatórios.

Para a análise com recuperação da informação inter-blocos tem-se:

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	-	-
Blocos Ajustados dentro Repet.	(k-1)r	Q_b	$\sigma^2 + k\sigma_b^2$	-
Tratamentos	t-1	Q_t	$\sigma^2[1+\sigma_b^2/(\sigma^2+k\sigma_b^2)] + r\sigma_t^2$	Q_t/Q_{EE}
Erro Efetivo	(k-1)(t-1)	Q_{EE}	$\sigma^2[1+\sigma_b^2/(\sigma^2+k\sigma_b^2)]$	-
Erro Intrabloco	(k-1) (t-1)	Q_{EI}	σ^2	-

Na análise com recuperação da informação inter-blocos, os blocos são ajustados para os efeitos de tratamentos para que seja obtida a estimativa do componente de variância entre blocos

(σ_b^2). Essa estimativa é usada na estimação das médias ajustadas de tratamentos. Nesse processo, o peso dado às informações de parcelas dentro de blocos é $1/\sigma^2$ e o peso dado às informações entre blocos é dado por $1/(\sigma^2 + k\sigma_b^2)$. Assim, as informações são ponderadas linearmente e aditivamente pelo inverso de suas precisões. Esse é o próprio princípio do procedimento BLUP, derivado sob o enfoque freqüentista ou bayesiano, o qual implicitamente pondera dessa forma as informações associadas a diferentes precisões. Também o procedimento REML para a estimativa de componentes de variância com dados desbalanceados baseia-se em princípio similar: pondera as informações de cada tratamento por uma função da quantidade de informação e precisão associada a cada um deles.

5.5 Arranjo Fatorial no Delineamento em Blocos ao Acaso

Modelo Estatístico

$Y_{ijk} = u + A_i + B_j + r_k + AB_{ij} + e_{ijk}$, em que Y_{ijk} é o valor medido no campo, u é a média geral, A_i é o efeito do tratamento i pertencente ao fator qualitativo A, B_j é o efeito do tratamento j pertencente ao fator qualitativo B, r_k é o efeito da repetição k, AB_{ij} é o efeito da interação do tratamento i pertencente ao fator qualitativo A com o tratamento j pertencente ao fator qualitativo B e e_{ijk} é o efeito do erro aleatório ou resíduo associado à cada parcela ijk.

a) Fatores repetições, tratamentos A e B e interação A x B de efeitos fixos (Modelo Fixo).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + [(ab)/(r-1)]\sum_i r_k^2$	Q_r/Q_e
Fator A	a-1	Q_a	$\sigma^2 + [(rb)/(a-1)]\sum_i A_i^2$	Q_a/Q_e
Fator B	b-1	Q_b	$\sigma^2 + [(ra)/(b-1)]\sum_j B_j^2$	Q_b/Q_e
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + [r/((a-1)(b-1))]\sum_{ij} AB_{ij}^2$	Q_{ab}/Q_e
Erro	(ab-1) (r-1)	Q_e	σ^2	-

Verifica-se que, no modelo completamente fixo, ambos os fatores e a interação são testados com o erro.

b) Fatores repetições, tratamentos A e B e interação A x B de efeitos aleatórios (Modelo aleatório).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + ab\sigma_r^2$	Q_r/Q_e
Fator A.	a-1	Q_a	$\sigma^2 + r\sigma_{ab}^2 + rb\sigma_a^2$	Q_a/Q_{ab}
Fator B	b-1	Q_b	$\sigma^2 + r\sigma_{ab}^2 + ra\sigma_b^2$	Q_b/Q_{ab}
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + r\sigma_{ab}^2$	Q_{ab}/Q_e
Erro	(ab-1) (r-1)	Q_e	σ^2	-

Verifica-se que, no modelo completamente aleatório, os fatores A e B são testados com a interação A x B e a interação é testada com o erro.

c) Fator tratamentos A de efeitos fixos, fator tratamentos B e interação A x B de efeitos aleatórios e fator repetições de efeitos aleatórios (Modelo Misto).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + ab\sigma_r^2$	Q_r/Q_e
Fator A	a-1	Q_a	$\sigma^2 + [ra/(a-1)]\sigma_{ab}^2 + + [(rb)/(a-1)]\sum_i A_i^2$	Q_a/Q_{ab}
Fator B	b-1	Q_b	$\sigma^2 + ra\sigma_b^2$	Q_b/Q_e
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + [ra/(a-1)]\sigma_{ab}^2$	Q_{ab}/Q_e
Erro	(ab-1) (r-1)	Q_e	σ^2	-

Verifica-se que, no modelo misto, o fator de efeitos fixos é testado com a interação e o fator de efeitos aleatórios e a interação são testados com o erro. A esperança de quadrados médio do fator B segue a abordagem adotada por Snedecor e Cochran e Steel e Torrie. Uma alternativa,

apresentada por Ridgman (1975), inclui o componente σ^2_{ab} nessa esperança de quadrado médio e, por conseqüência, o fator B é testado com o quadrado médio da interação A x B. Essa última abordagem produz componentes de variância idênticos aos que se obtém pela metodologia de modelos mistos via o procedimento REML.

5.6 Arranjo em Parcela Subdividida no Delineamento em Blocos ao Acaso

Modelo Estatístico

$Y_{ijk} = u + A_i + B_j + r_k + rA_{ik} + AB_{ij} + e_{ijk}$, em que Y_{ijk} é o valor medido no campo, u é a média geral, A_i é o efeito do tratamento i pertencente ao fator qualitativo A alocado nas parcelas principais, B_j é o efeito do tratamento j pertencente ao fator qualitativo B alocado nas sub-parcelas, r_k é o efeito da repetição k, rA_{ik} é o efeito da interação do fator A com repetições ou erro (a) associado às parcelas principais, AB_{ij} é o efeito da interação do tratamento i pertencente ao fator qualitativo A com o tratamento j pertencente ao fator qualitativo B e e_{ijk} é o efeito do erro aleatório associado à cada parcela ijk ou erro (b).

a) Fatores tratamentos A e B e interação A x B de efeitos fixos e fator repetições de efeitos aleatórios (Modelo fixo, exceto repetições).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + b\sigma^2_e + ab\sigma^2_r$	Q_r/Q_{Ea}
Fator A	a-1	Q_a	$\sigma^2 + b\sigma^2_e + + [(rb)/(a-1)]\sum_i A_i^2$	Q_a/Q_{Ea}
Erro (a)	(r-1)(a-1)	Q_{Ea}	$\sigma^2 + b\sigma^2_e$	-
Fator B	b-1	Q_b	$\sigma^2 + [(ra)/(b-1)]\sum_j B_j^2$	Q_b/Q_{Eb}
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + [r/((a-1)(b-1))]\sum_{ij} AB_{ij}^2$	Q_{ab}/Q_{Eb}
Erro (b)	a(b-1) (r-1)	Q_{Eb}	σ^2	-

Verifica-se que, no modelo completamente fixo, o fator A nas parcelas é testado com o erro (a) e o fator B nas subparcelas e a interação são testados com o erro (b).

b) Fatores repetições, tratamentos A e B e interação A x B de efeitos aleatórios (Modelo aleatório).

Fontes de Variação	**GL**	**QM**	**E(QM)**	**F sob H_o**
Repetições	r-1	Q_r	$\sigma^2 + b\sigma_e^2 + ab\sigma_r^2$	Q_r/Q_{Ea}
Fator A.	a-1	Q_a	$\sigma^2 + b\sigma_e^2 + r\sigma_{ab}^2 + rb\sigma_a^2$	$(Q_a + Q_{Eb})$ / $(Q_{ab} + Q_{Ea})$
Erro (a)	(r-1)(a-1)	Q_{Ea}	$\sigma^2 + b\sigma_e^2$	-
Fator B	b-1	Q_b	$\sigma^2 + r\sigma_{ab}^2 + ra\sigma_b^2$	Q_b/Q_{ab}
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + r\sigma_{ab}^2$	Q_{ab}/Q_{Eb}
Erro (b)	a(b-1) (r-1)	Q_{Eb}	σ^2	-

Verifica-se que, no modelo completamente aleatório, o fator A nas parcelas é testado com o erro (a) e com a interação, o fator B nas subparcelas é testado com a interação e a interação é testada com o erro (b).

c) Fator tratamentos A de efeitos fixos, fator tratamentos B e interação A x B de efeitos aleatórios e fator repetições de efeitos aleatórios (Modelo Misto).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + b\sigma_e^2 + ab\sigma_r^2$	Q_r/Q_{Ea}
Fator A	a-1	Q_a	$\sigma^2 + b\sigma_e^2 + [ra/(a-1)]\sigma_{ab}^2 + + [(rb)/(a-1)]\sum_i A_i^2$	$(Q_a + Q_{Eb})$ / $(Q_{ab} + Q_{Ea})$
Erro (a)	(r-1)(a-1)	Q_{Ea}	$\sigma^2 + b\sigma_e^2$	-
Fator B	b-1	Q_b	$\sigma^2 + ra\sigma_b^2$	Q_b/Q_{Eb}
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + [ra/(a-1)]\sigma_{ab}^2$	Q_{ab}/Q_{Eb}
Erro (b)	a(b-1) (r-1)	Q_{Eb}	σ^2	-

Verifica-se que, nesse modelo misto, o fator A (de efeitos fixos) nas parcelas é testado com o erro (a) e com a interação e o fator B de efeitos aleatórios e a interação são testados com o erro (b).

d) Fator tratamentos A de efeitos aleatórios, fator tratamentos B de efeitos fixos e interação A x B e fator repetições de efeitos aleatórios (Modelo Misto).

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Repetições	r-1	Q_r	$\sigma^2 + b\sigma_e^2 + ab\sigma_r^2$	Q_r/Q_{Ea}
Fator A	a-1	Q_a	$\sigma^2 + b\sigma_e^2 + rb\sigma_a^2$	Q_a/Q_{Ea}
Erro (a)	(r-1)(a-1)	Q_{Ea}	$\sigma^2 + b\sigma_e^2$	-
Fator B	b-1	Q_b	$\sigma^2 + [rb/(b-1)]\sigma_{ab}^2 + + [(ra)/(b-1)]\sum_i B_j^2$	Q_b/Q_{ab}
Interação A x B	(a-1)(b-1)	Q_{ab}	$\sigma^2 + [rb/(b-1)]\sigma_{ab}^2$	Q_{ab}/Q_{Eb}
Erro (b)	a(b-1) (r-1)	Q_{Eb}	σ^2	-

Verifica-se que, nesse outro modelo misto, o fator A (de efeitos aleatórios) nas parcelas é testado com o erro (a), fator B de efeitos fixos é testado com a interação e a interação é testada com o erro (b).

5.7 Arranjo Hierárquico no Delineamento em Blocos ao Acaso

Modelo Estatístico

$Y_{ijk} = u + g_i + s_{ij} + b_k + e_{ijk}$, em que Y_{ijk} é o valor medido no campo, u é a média geral, g_i é o efeito do grupo i, s_{ij} é o efeito do subgrupo j dentro do grupo i, b_k é o efeito do bloco k e $e_{(i)jk}$ é o efeito do erro aleatório ou resíduo associado à parcela ijk.

a) Fatores grupo, subgrupo e blocos de efeitos aleatórios.

Fontes de Variação	GL	QM	E(QM)	F sob H_o
Blocos	b-1	Q_b	-	-
Fator Grupo	g-1	Q_g	$\sigma^2 + b\sigma_s^2 + bs\sigma_g^2$	Q_g/Q_s
Fator Subgrupo dentro Grupo	(s-1)g	Q_s	$\sigma^2 + b\sigma_s^2$	Q_s/Q_r
Resíduo	(b-1)(qs-1)	Q_r	σ^2	-

Nesse caso, os efeitos de grupos são testados com os efeitos de subgrupos e esses são testados com os resíduos.

6 TESTES DE HIPÓTESES E COMPARAÇÕES MÚLTIPLAS

Considerando um modelo mais simples, como aquele associado à avaliação de tratamentos de efeitos fixos em um experimento no delineamento inteiramente casualizado, tem-se que:

$Y_{ij} = \mu + t_i + e_{ij}$

Y_{ij} : valor observado do i-ésimo tratamento na repetição j.

μ : efeito da média geral, fixo, $E(\mu) = \mu$ e $E(\mu^2) = \mu^2$.

t_i : efeito do tratamento i, fixo, $E(t_i) = t_i$ e $E(\mathrm{t}_\mathrm{i}^2) = \mathrm{t}_\mathrm{i}^2$.

e_{ij}: erro aleatório, $E(e_{ij}) = 0$ e $E(\mathrm{e}_\mathrm{ij}^2) = \sigma^2$, ou seja,

$e_{ij} \sim N(0,\ \sigma^2)$, de forma que os erros apresentam distribuição normal, são independentes e identicamente distribuídos com variância comum σ^2.

Na especificação do modelo, verifica-se:

(i) O modelo é aditivo nos parâmetros, ou seja, os efeitos da média geral, de tratamentos e do erro são aditivos e não correlacionados.

(ii) Os erros apresentam distribuição normal.

(iii) Os erros são independentes, ou seja, não apresentam correlação serial.

(iv) Os erros são homocedásticos, ou seja, apresentam homogeneidade de variâncias.

Tal modelo associa-se a uma análise de variância do tipo:

FV	GL	QM	E(QM)	F
Tratamentos	t-1	Q_1	$\sigma^2 + \frac{r}{t-1}\sum_{i=1}^{t} t_i^2$	Q_1/Q_2
Erro	(r-1) t	Q_2	σ^2	
Total	rt-1			

Considerando que uma razão entre variâncias tem distribuição F de Snedecor, o teste F, em princípio, destina-se à comparação entre variâncias. No entanto, presta-se também à comparação de $\sigma^2 + \frac{r}{t-1}\sum_{i=1}^{t} t_i^2$ com σ^2, ou seja, no teste da igualdade de Q_1 com Q_2. Neste caso, se for comprovado que $Q_1 = Q_2$, tem-se $\sum_{i=1}^{t} t_i^2$ estatisticamente igual a zero, fato que permite concluir que os dados

podem ser tratados como uma única amostra de rt itens provenientes de $Y \sim N\ (\mu,\ \sigma^2)$ ou, em termos práticos, que não existem diferenças reais entre os tratamentos.

Formalmente, é necessário formular as hipóteses estatísticas, que podem ser denominadas: (i) **H_o: hipótese de nulidade** ou da não diferença entre tratamentos; (ii) **H_a: hipótese alternativa,** hipótese de pesquisa ou do pesquisador ou hipótese da diferença entre tratamentos. A hipótese H_o pode ser verdadeira ou falsa, e sua aceitação ou rejeição por parte do pesquisador está sempre associada a um grau de incerteza na tomada da decisão sobre tal hipótese estatística.

Associada a esta decisão, o pesquisador pode cometer os erros tipo I e tipo II, conforme ilustrado a seguir.

Conclusão do Teste Estatístico	**Situação real na população**	
	H_o verdadeira	**H_o falsa**
Rejeitar H_o	Erro Tipo I $\rightarrow \alpha$	Decisão Correta $\rightarrow$ (1-β)
Aceitar H_o	Decisão Correta $\rightarrow$ (1-α)	Erro Tipo II $\rightarrow \beta$

Assim, o pesquisador incorre no erro tipo I, quando rejeita uma hipótese H_o que é verdadeira e incorre no erro tipo II, quando aceita uma hipótese H_o que é falsa. A probabilidade de cometer um erro tipo I é designada por α e o maior valor de α para H_o verdadeira é denominado **nível de significância** de um teste estatístico, ou seja, a significância de um teste é a probabilidade máxima que se admite correr o risco de cometer um erro tipo I. A probabilidade (1-α) é denominada **grau de confiança.** A probabilidade de cometer um erro tipo II é designada por β. A probabilidade (1-β) é denominada **poder** de um teste e refere-se à probabilidade de se rejeitar H_o, quando H_o é falsa, ou seja, à capacidade de detectar as diferenças reais que existem entre tratamentos.

Na prática, o pesquisador tem controle apenas sobre o tipo I já que pode escolher, nominalmente, o nível de significância α. Quanto ao erro tipo II, não há como controlá-lo. No caso do teste F, pode ser estabelecida a hipótese:

$H_o = \sum_{i=1}^{t} t_i = 0$, ou seja, os efeitos de tratamentos não diferem entre si.

Caso esta hipótese seja rejeitada a um nível de significância α, conclui-se que existe pelo menos um contraste significativo indicando diferença entre duas médias, pelos testes t de Student e F de Snedecor. Sendo a média de cada tratamento estimada por $\overline{Y}_{i.} = \sum_{j=1}^{r} Y_{ij} / r$ e a variância da média de tratamento, por $\hat{V}ar\ (\overline{Y}_{i.}) = \hat{\sigma}^2 / r$ (em que $\hat{\sigma}^2 = Q_2$), pode-se comparar todas as médias de tratamentos mediante procedimentos de comparações múltiplas, os quais fazem uso de $\overline{Y}_{i.}$ e de $(\hat{\sigma}^2 / r)$.

Existem vários procedimentos para comparações múltiplas (Steel e Torrie, 1980; Neter et al., 1996). Dentre as características desejáveis destes procedimentos, sobressaem a baixa probabilidade de erro tipo I e o alto poder dos testes, ambos os conceitos associados à rejeição da hipótese H_o. O controle do erro tipo I pode ser realizado por comparação (a probabilidade de erro tipo I é dada pela razão entre o número de erros tipo I, em comparações duas a duas e o número total de comparações envolvendo todas as combinações duas a duas entre tratamentos em todos os experimentos) ou por experimento (a probabilidade de erro tipo I é dada pela razão entre o número de experimentos com pelo menos um erro tipo I e o número total de experimentos). Segundo Perecin e Barbosa (1988), a questão prática é que um teste com taxa de erro por comparação torna-se muito frouxo ao ser aplicado em todo o experimento, enquanto que um teste com taxa de erro por experimento torna-se conservador ou rigoroso ao ser olhado por comparação. Uma explicação simples e acessível é dada por Vieira (1999).

Os principais testes de comparações múltiplas são: t (ou LSD de Fisher ou forma F protegida de Fisher, quando aplicado somente quando o teste F é significativo), Tukey (ou HSD), Duncan, Scheffé, Dunnet, Newman-Keuls (ou Student-Newman-Keuls – SNK ou simplesmente Keuls), Bonferroni (ou LSDB) e Scott-Knott. Destes, o teste de Dunnet é adequado para comparação das médias dos tratamentos com as médias das testemunhas e os demais permitem comparações entre as médias dos vários tratamentos. É importante destacar que todos estes testes foram derivados na suposição de tratamentos de efeitos fixos e, portanto, não se justifica a sua aplicação quando os efeitos de tratamentos são considerados aleatórios.

Os testes t não protegido e Duncan controlam as taxas de erro tipo I por comparação, ao passo que os testes de Tukey e Scheffé controlam as taxas de erro tipo I, por experimento (Chew,

1986). Assim, ao serem olhados por comparação, os testes menos rigorosos são o t e o Duncan e os mais rigorosos são o Tukey e o Scheffé. O teste de Newman-Keuls controla a taxa de erro por comparação mas não controla totalmente a taxa de erro por experimento, apresentando, portanto, um rigor intermediário. Assim, dos menos rigorosos para os mais rigorosos têm-se: t, Duncan, Newman-Keuls, Tukey, Scheffé. O rigor, no caso, refere-se à maior dificuldade em detectar diferenças significativas entre tratamentos. A um maior rigor está associado um menor poder. Como o nível de significância e o poder de um teste crescem juntos, uma alternativa para aumentar o poder de testes muito rigorosos é aumentar o nível de significância, por exemplo, adotando α = 10%.

O teste de Bonferroni refere-se a uma modificação do teste t não protegido, visando conservar a taxa de erro tipo I, por experimento. Neste caso, o rigor em relação ao teste t é aumentado. O teste de Scott-Knott controla bem a taxa de erro tipo I, por experimento e por comparação, conservando-a em torno ou abaixo do nível nominal de significância (Silva et al., 1999).

Outro teste cujo uso é freqüente é o t bayesiano ou teste de Waller-Duncan. Este teste não envolve o conceito usual de taxas de erro e também utiliza o valor calculado da estatística F no cálculo do valor crítico de t*. Se o valor de F calculado é grande (indicando tratamentos heterogêneos), o valor crítico é reduzido, tornando o teste mais poderoso. Se o valor calculado de F é pequeno, o valor crítico de t* será aumentado, tornando o teste menos poderoso, ou seja, mais rigoroso (Chew, 1986).

Uma avaliação (via simulação) de vários testes quanto às taxas de erro tipo I e poder foi realizada por Perecin e Barbosa (1988) e revelou os seguintes resultados:

- Os testes t e Duncan não apresentam taxas favoráveis de erro tipo I, ou seja, não conservam adequadamente os níveis nominais de significância;
- O teste de Tukey possui um poder muito reduzido, especialmente quando há grande número de tratamentos;
- O teste t bayesiano concilia de certa forma as características de alto poder e baixas taxas de erro tipo I, entretanto, as taxas de erro tipo I não podem ser previstas com exatidão e dependem do número de tratamentos e de suas magnitudes, requerendo, portanto, aplicação cuidadosa;

- O teste de Newman-Keuls possui poder muito superior e taxas de erro tipo I similares ao teste de Tukey.

Assim, os testes t, Duncan e Tukey, amplamente utilizados no Brasil, não são os mais recomendados e só devem ser usados com ressalvas. O teste de Newman-Keuls, pouco utilizado no Brasil, é altamente recomendado tendo em vista as taxas favoráveis de erro tipo I, o relativo alto poder e o rigor intermediário. Dessa forma, pode ser utilizado sem maiores cuidados. Em realidade, tal teste vem sendo amplamente utilizado (em detrimento dos demais) pelos franceses e também nos trabalhos de melhoramento de plantas perenes em países da África.

A título de exemplo, será considerada a aplicação do teste de Newman-Keuls. Este teste usa a amplitude estudentizada ou padronizada (q) (assim como o fazem os testes Tukey e Duncan) e é decorrente dos trabalhos de Newman (1939) e Keuls (1952).

Todos os testes baseiam-se na especificação de uma diferença mínima significativa (DMS ou LSD) entre as médias de tratamentos, a qual está associada a um nível de significância α. Esta DMS, em conjunto com uma **estatística de teste** dada pelo contraste de médias de dois tratamentos, gera uma **regra de decisão.** A DMS permite testar a hipótese $H_o : (\mu + t_i) - (\mu + t_i^{'}) = 0$, ou seja, a hipótese da nulidade da diferença das médias de tratamentos. Se o contraste de médias for superior à DMS, rejeita-se a hipótese H_o e os tratamentos são tomados como diferentes estatisticamente. Esta é a **regra de decisão.**

Para o teste de Newman-Keuls, esta DMS é dada por:

$DMS = q_{(m,\eta,\alpha)}[\hat{\sigma}/(r)^{1/2}] = q_{(m,\eta,\alpha)}[Q_2/r]^{1/2}$, em que:

$q_{(m,\eta,\alpha)}$: valor da amplitude padronizada (que pode ser obtida de uma tabela para o teste de Tukey) associada ao número m de médias ordenadas abrangidas pela comparação, ao número η de graus de liberdade do resíduo da análise de variância e nível de significância α (1 %, 5 % ou 10 %).

$[Q_2/r]^{1/2} = \hat{\sigma}/(r)^{1/2}$: erro padrão da média de tratamentos, função do quadrado médio do resíduo (Q_2) da análise de variância.

Esta expressão é similar às expressões da DMS empregadas pelos testes de Duncan e de Tukey. Difere da expressão do teste de Duncan devido ao fato de o nível de significância α não ser constante no teste de Duncan (este também é um ponto desfavorável deste teste) e ser constante para o teste de Newman-Keuls. Difere do teste de Tukey em razão de o valor m ser constante no teste de Tukey e variável no teste de Newman-Keuls, de acordo com o número de médias abrangidas pela comparação ou contraste.

Considerando os dados apresentados por Steel e Torrie (1980), têm-se as seguintes médias para seis tratamentos: T_1 = 13,3; T_2 = 14,6; T_3 = 18,7; T_4 = 19,9; T_5 = 24,0 e T_6 = 28,8. Sendo o quadrado médio do resíduo Q_2 = 11,79, r = 5 repetições e η = 24, têm-se os seguintes valores de $q_{(m, 24, 0,05)}$ e de DMS para um nível de significância α = 5 %:

m	$q_{(m, 24, 0,05)}$	DMS
2	2,92	4,5
3	3,53	5,4
4	3,90	6,0
5	4,17	6,4
6	4,37	6,7

Comparando-se inicialmente T_1 com T_6, tem-se T_6-T_1 = 15,5 e DMS_6 = 6,7 (esta é a DMS única que seria usada no teste de Tukey), de forma que T_6 é significativamente diferente de T_1. As próximas comparações a serem realizadas são, então, T_1 com T_5 e T_2 com T_6. Sendo $T_5 – T_1$ = 10,7 e $T_6 – T_2$ = 14,2 e a DMS_5 = 6,4, conclui-se que $T_1 \neq T_5$ e $T_2 \neq T_6$. As próximas comparações a serem feitas são T_4 com T_1, T_5 com T_2 e T_6 com T_3. Sendo $T_4 – T_1$ = 6,5 , $T_5 – T_2$ = 9,4 e $T_6 – T_3$ = 10,1 e a DMS_4 = 6,0, conclui-se que $T_4 \neq T_1$, $T_5 \neq T_2$ e $T_6 \neq T_3$. Em seguida, faz-se as comparações de T_3 com T_1 , T_4 com T_2 , T_5 com T_3 e de T_6 com T_4, donde concluiu-se que $T_1 = T_3$, $T_4 = T_2$, $T_5 = T_3$ e $T_6 \neq T_4$, uma vez que $T_3 – T_1$ = 5,4, $T_4 – T_2$ = 5,3, $T_5 – T_3$ = 5,3, $T_6 – T_4$ = 8,9 e DMS_3 = 5,4.

Neste ponto concluiu-se que $T_1 = T_2$, $T_2 = T_3$, $T_4 = T_3$, $T_5 = T_4$, não sendo necessário testar os referidos contrastes. Entretanto, falta contrastar ainda T_5 com T_6, visto que concluiu-se que $T_6 \neq T_4$. Sendo $T_6 – T_5$ = 4,8 e DMS_2 = 4,5, concluiu-se que $T_5 \neq T_6$. Em resumo e considerando letras iguais unindo médias estatisticamente não diferentes, tem-se:

T_1	T_2	T_3	T_4	T_5	T_6
13,3 d	14,6 cd	18,7 bcd	19,9 bc	24,0 b	28,0 a

Verifica-se que tal teste conduziu a ambigüidade nos resultados, conforme mostrado pelas sobreposições das letras que unem os tratamentos estatisticamente não diferentes. Isto ocorre também para os demais testes, exceto para o de Scott e Knott (1974). Tal teste é baseado em uma análise de agrupamento, a qual conduz a grupos mutuamente exclusivos de tratamentos e, portanto, elimina o problema de ambigüidade. Além desta propriedade desejável, o teste de Scott-Knott também é favorável em termos das taxas de erro tipo I e apresenta poder superior aos demais testes (Silva et al., 1999), sendo por isto fortemente recomendado. Ainda segundo Silva et al. (1999), este teste é conservador ou rigoroso mas não tanto quanto os testes de Scheffé, Tukey e Newman-Keuls.

No caso de tratamentos de efeitos aleatórios, os procedimentos tradicionais de comparação múltipla não se aplicam pois foram derivados com suposição de tratamentos de efeitos fixos. Mas os efeitos aleatórios preditos, em conjunto com a estimativa desvio padrão do erro de predição (SEP) podem ser usados para a obtenção de intervalos de confiança dos efeitos aleatórios preditos por meio da expressão $(u + t_i) \pm t\ SEP$, em que t = 1.96 é o valor tabelado da distribuição t de Student associada a um grande número de graus de liberdade. Verificando-se a sobreposição desses intervalos de confiança pode-se inferir, de forma aproximada, sobre comparações múltiplas entre efeitos aleatórios de tratamentos. Detalhes sobre testes de hipóteses de efeitos aleatórios são apresentados no Capítulo 4.

7 DELINEAMENTOS EXPERIMENTAIS ÓTIMOS NO CASO DE TRATAMENTOS CORRELACIONADOS E ERROS CORRELACIONADOS

Os delineamentos experimentais tradicionais abordados anteriormente foram otimizados e são ótimos tendo-se por base as seguintes condições:

(i) os tratamentos são de efeitos fixos.

(ii) as estimativas dos efeitos de tratamentos são obtidas por estimadores pertencentes à classe melhores estimadores lineares não viciados (BLUE).

(iii) os tratamentos são não correlacionados.

(iv) Os erros são não correlacionados.

Muitas das situações reais de experimentação não atendem a essas condições. Com erros espacialmente correlacionados, delineamentos ótimos podem ser obtidos levando-se em consideração a estrutura de correlação e os métodos e modelos de análise espacial a serem aplicados no processo de estimação ou predição dos efeitos de tratamentos. Delineamentos ótimos nesse sentido foram apresentados por Cullis et al. (2006).

Na situação de tratamentos de efeitos aleatórios e correlacionados, tal como ocorre na experimentação em melhoramento genético, devido ao parentesco entre os tratamentos em avaliação, nenhuma garantia existe que esses delineamentos tradicionais são realmente ótimos. Ótimos aqui designa os melhores possíveis. Essa dúvida existe sobretudo porque outro procedimento estatístico (BLUP) é usado para a predição dos efeitos de tratamentos e conseqüentemente para as inferências.

Para grande número de tratamentos a serem avaliados, os delineamentos alfa são tidos como ótimos. Esses delineamentos pertencem à classe dos delineamentos em blocos incompletos resolvíveis (caso em que os blocos podem ser agrupados em uma repetição completa contendo todos os tratamentos, permitindo, se for o caso, que seja realizada análise em blocos completos) e foram otimizados com base na abordagem tradicional de efeitos fixos e estimativas BLUE para os efeitos de tratamentos não correlacionados. Visando verificar a optimalidade desses delineamentos para o caso de tratamentos correlacionados, efeitos aleatórios e predições BLUP para os efeitos de

tratamentos, Piepho e Williams (2006) conduziram um estudo de simulação, comparando-os com o delineamento em parcelas subdivididas agrupando os tratamentos aparentados como subparcelas de um grupo. Concluíram que o delineamento alfa com casualização completa é superior ao delineamento em parcela subdividida em termos da predição BLUP e também da estimação BLUE dos efeitos de tratamentos. Concluíram também, para o delineamento alfa, que o procedimento BLUP foi superior ao BLUE em termos de habilidade preditiva dos verdadeiros efeitos de tratamentos.

Entretanto, para o delineamento em parcela subdividida, o procedimento BLUE mostrou-se superior ao BLUP, corroborando o fato de tal delineamento ter sido otimizado sob estimação BLUE de efeitos fixos. O delineamento alfa também foi otimizado dessa forma, mas nele os efeitos de blocos são considerados de efeitos aleatórios para recuperação de informação interblocos. E essa recuperação de informação é também uma característica do BLUP, de forma que menor diferença entre BLUP e BLUE é esperada nesse delineamento.

O delineamento com estrutura de parcela subdividida enfatiza o poder na comparação entre tratamentos dentro de grupo. E o interesse do melhoramento é na comparação entre tratamentos no geral. Bueno Filho e Gilmour (2003) consideraram também a questão de delineamentos ótimos com tratamentos correlacionados e de efeitos aleatórios e relatam que o delineamento ótimo depende da estrutura de correlação genética entre os tratamentos e também da herdabilidade. Assim, em função dessas informações, é possível definir melhor quais tratamentos devem participar de quais blocos, visando maximizar a eficiência do delineamento. Em outras palavras, é possível encontrar um delineamento ótimo, na classe dos blocos incompletos resolvíveis, para o caso de tratamentos de efeitos aleatórios e correlacionados a serem preditos via BLUP. O delineamento alfa enfatiza controle local em apenas uma direção. Por controlar a heterogeneidade ambiental em duas direções, os delineamentos em linha e coluna podem ser mais eficientes do que os alfa.

CAPÍTULO 2
ANÁLISE EXPLORATÓRIA DE DADOS

1 QUALIDADE DOS DADOS

Além do adequado planejamento e delineamento dos experimentos de campo, conforme visto no Capítulo 1, inferências fidedignas a partir desses ensaios científicos dependem essencialmente da qualidade dos dados, da adoção do modelo matemático ou estatístico plausível, da adoção de procedimentos adequados de estimação/predição, da disponibilidade de algoritmos e programas computacionais eficientes. Destes quatro requisitos essenciais, a qualidade dos dados deve ser avaliada primeiro, pois, com uma base de dados inadequada, os demais requisitos serão completamente prejudicados.

Como qualidade dos dados, ressaltam-se os aspectos: presença de erros de mensuração e de digitação; presença de dados atípicos (*outliers*); conjunto de dados que não atendem aos pressupostos dos modelos de análise. Esses pressupostos são: (i) para os modelos tradicionais de análise de variância e de regressão (aditividade, homocedasticia, normalidade e independência de

erros); (ii) para a estimação por máxima verossimilhança e máxima verossimilhança restrita (REML): normalidade. A estimação por REML pode acomodar variâncias heterogêneas e dependência entre erros. Igualmente, modelos multiplicativos podem ser ajustados via REML.

Os conjuntos de dados de experimentos de campo podem apresentar certa porcentagem de dados defeituosos causados por danos físicos. Isto é comum, por exemplo, em experimentos com espécies florestais, que apresentam certa fração de árvores com copa ou tronco quebrado, que reduzem artificialmente o crescimento das árvores. A inclusão deste tipo de observação nas análises pode tornar imprecisos os resultados, visto que a ocorrência de tais dados pode diferir aleatoriamente entre os tratamentos. Um ajuste dos dados via covariância é freqüentemente impreciso, devido às várias causas dos danos. Assim, observações defeituosas devem ser descartadas dos conjuntos de dados, para efeito de análise estatística. Posteriormente à análise, as inferências práticas (resultados) podem ser ajustadas considerando uma fração média de ocorrência de tal dano físico no ambiente alvo da inferência.

As observações atípicas ou *outliers* ou pontos de influência são definidos como dados que desviam significativamente de suas esperanças matemáticas segundo uma distribuição de probabilidade (Hawkins, 1980). De acordo com Hoaglin et al. (1983), as coleções reais de dados apresentam em geral cerca de 5 % de *outliers*, sendo que sob o modelo Gaussiano, o esperado seria menos de 1 %. Estas observações atípicas podem impactar, por exemplo, médias e variâncias, causando inflação das variâncias estimadas e, por conseqüência, conduzindo a uma inferência estatística pobre. As médias tendem a ser robustas quando o número de *outliers* é pequeno ($\leq$ 5 %).

O passo inicial de uma análise estatística deve enfatizar a identificação de possíveis *outliers*. Várias técnicas gráficas, medidas de pontos de influência e testes podem ser aplicados para tal fim no contexto da análise de resíduos. Um meio simples de identificá-los refere-se aos gráficos de resíduos padronizados versus valores preditos. Se o valor absoluto de qualquer resíduo padronizado é >2,5, a observação pode ser tratada como um *outlier* (Fernandez, 1992). Uma abordagem similar é sugerida por Tukey (1977), a qual indica a exclusão de dados com valores superiores a 2,7 desvios padrões (2,7 vezes a raiz quadrada da variância fenotípica individual). No entanto, é importante ter em mente que o objetivo é eliminar apenas dados aberrantes e que se tornaram assim devido a

causas não aceitáveis. A detecção de outlier via modelagem usando REML foi proposta por Thompson (1985).

Após a identificação de *outliers* e sua proporção, técnicas robustas de estimação de parâmetros (médias e variâncias, por exemplo) podem ser adotadas. Técnicas robustas ou resistentes de estimação produzem estimativas insensíveis à deleção e/ou inclusão de um pequeno número de observações. Robustez significa também procedimentos estatísticos livres de suposições não realistas. Assim, de maneira genérica, técnicas podem ser robustas a dois fatores: presença de *outliers* e incorreta especificação do modelo estatístico. No contexto do primeiro caso, estimativas robustas podem ser obtidas: via um esquema de ponderação, a qual dá peso às observações de acordo com uma probabilidade pré-definida de obtenção de uma observação, segundo uma distribuição de probabilidade assumida; via uma censura balanceada dos dados; eliminando as caudas da distribuição dos dados; por meio do uso de estimadores quantais; por meio do uso de distribuições com cauda pesada.

Em geral, a estimação robusta é preferível em relação à meramente conduzir a análise sem os *outliers*. Os métodos robustos dão muita ênfase à porção principal dos dados e pouca aos *outliers*, fato que é uma propriedade atrativa.

A qualidade de dados refere-se, também, em essência, à quantidade de dados. Segundo Hoaglin et al. (1983), a estatística clássica sublima e apoia-se em argumentos de consistência, variância assintótica e eficiência assintótica, reportando-se a amostras de grande dimensão. Entretanto, na prática, os tamanhos amostrais são pequenos, fato que demanda um exame criterioso das técnicas estatísticas nesta situação. Tamanhos amostrais para estimação e predição em várias situações práticas no melhoramento genético são relatados por Resende (2002). No Capítulo 1 são também apresentados aspectos referentes ao tamanho amostral.

Outro fator a considerar é o desbalanceamento dos dados. Na atualidade, desbalanceamento não significa dados problemáticos visto que os métodos de estimação e predição disponíveis permitem lidar adequadamente com este fator.

2 DISTRIBUIÇÕES DE PROBABILIDADE E SUAS CARACTERÍSTICAS

O conhecimento da distribuição de probabilidade associada ao conjunto de dados a ser analisado é relevante, pois, além de permitir uma melhor caracterização da estrutura dos dados, permite verificar a adequacidade dos métodos de análise (estimação, predição, inferência) a serem aplicados. Assim, pode-se direcionar o tipo de transformação (se necessário) dos dados antes da análise, bem como os procedimentos estatísticos a serem empregados.

Os conjuntos de dados contêm uma ou mais variáveis denominadas variáveis aleatórias. Variável aleatória refere-se a uma regra de associação de um valor numérico a cada ponto do espaço amostral. Formalmente, tem-se que uma variável aleatória Y em um espaço de probabilidade (Ω, σ-A, P) é uma função real definida em Ω, tal que o evento [Y$\leq$y] é aleatório $\forall y \in \Re$, ou seja, a função Y: $\Omega \rightarrow \Re$ é variável aleatória se o evento [Y$\leq$y] $\in \sigma$-A, $\forall y \in \sigma$-A. As denominações Ω, σ-A e P referem-se ao espaço amostral de um experimento, a sigma-álgebra de Ω e a medida de probabilidade, respectivamente.

A distribuição de probabilidade da variável aleatória Y é caracterizada pelos próprios valores de Y e pela regra ou função que associa, a cada valor (no caso de uma distribuição discreta), uma probabilidade. Esta função é denominada função de probabilidade no caso de variáveis aleatórias discretas e função densidade de probabilidade no caso de variáveis aleatórias contínuas. Uma variável aleatória é discreta quando o seu domínio é um conjunto finito ou infinito enumerável. Quando o domínio da variável aleatória é um conjunto infinito, tal é denominada contínua.

2.1 Distribuições Discretas

Uma distribuição discreta de probabilidade de uma variável aleatória Y é um conjunto de valores y de Y juntamente com suas probabilidades associadas. A função P (Y=y) é uma função de probabilidade desde que satisfaça às condições:

$$\text{(i)} \quad P(Y=y) \geq 0 \qquad \text{(ii)} \quad \sum P(Y=y) = 1$$

(a) Distribuição Bernoulli

Assume os valores 0 ou 1, de acordo com a presença ou ausência de um atributo. Como exemplo, tem-se: classificação de peças em perfeitas e defeituosas; classificação de sementes em duas cores distintas; classificação de plantas em vivas ou mortas.

Uma variável com distribuição Bernoulli com parâmetro θ tem função de probabilidade:

$P(Y = y) = \theta^y (1-\theta)^{1-y}$, $y = 0,1$, $0 < \theta < 1$

Assim, $P(Y = 1) = \theta$, $P(Y = 0) = 1-\theta$ e $\sum P(Y = y) = 1$

Média: θ

Variância: $\theta (1-\theta)$

(b) Distribuição Binomial

A distribuição Binomial conta o número de sucessos em n provas do tipo Bernoulli. Uma variável aleatória tem distribuição Binomial com parâmetros n e θ quando assume valores no conjunto (0, 1, 2, ..., n) e sua função de probabilidade é dada por:

$$P\ (Y = y) = \binom{n}{y} \theta^y\ (1-\theta)^{n-y},$$

$$y = 0,\ 1,\ 2,\ n,\ 0 < \theta < 1$$

Média: $n\theta$

Variância: $n\theta (1-\theta)$

A variável ou proporção $\hat{\theta} = \frac{y}{n}$, onde y~b (n, θ) tem como média $E(\hat{\theta}) = \theta$ e como variância $V(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$.

(c) Distribuição Poisson

Em geral, eventos raros como contagens por unidade de tempo e espaço obedecem a uma distribuição Poisson. A função de probabilidade de uma variável aleatória com distribuição Poisson com parâmetro θ é dada por:

$$P\ (Y = y) = \frac{\theta^{y} \cdot e^{-\theta}}{y!},\ \ y = 0,\ 1,\ 2,\ 3\ ...;\ \ \theta > 0,\ \ e = 2{,}71828$$

Média: θ

Variância: θ

A variável ou proporção $\hat{\theta} = \frac{y}{n}$, onde y~P(θ), tem como média $E(\hat{\theta}) = \theta$ e como variância $V(\hat{\theta}) = \theta / n$.

Outras distribuições discretas relevantes e suas características são apresentadas na Tabela 3, conforme Mood et al. 1974.

2.2 Distribuições Contínuas

Uma variável aleatória contínua não possui uma função de probabilidade que associe probabilidades a cada ponto ou valores de seu domínio. Estas probabilidades são calculadas para intervalos de valores do domínio através de uma função densidade de probabilidade. A função f(Y) é uma função densidade de probabilidade desde que satisfaça às condições:

(i) $P\ (a < Y < b) = \int_{a}^{b} f(y)\ dy$ (ii) $\int_{-\infty}^{\infty} f(y)\ dy = 1$

(a) Distribuição Normal

Uma variável com distribuição Normal ou Gaussiana com parâmetros μ e σ^2, tem como função densidade de probabilidade:

$$f(y) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y-u)^2\right\},\ y \in \Re,\ \mu \in \Re \text{ e } \sigma > 0$$

Média: μ

Variância: σ^2

Tabela 3. Distribuições discretas de probabilidade.

Distribuição	Função de Probabilidade	Espaço Paramétrico	Média $\mu = E(Y)$	Variância $\sigma^2 = E[(Y-\mu)^2]$
Uniforme Discreta	$f(y) = \frac{1}{N} I_{(1, \ldots, N)}(y)$	N =1, 2 , ...	$\frac{N+1}{2}$	$\frac{N^2-1}{12}$
Bernoulli	$f(y) = p^y q^{1-y} I_{(0, 1)}(y)$	$0 \le p \le 1$ $(q = 1-p)$	p	pq
Binomial	$f(y) = \binom{n}{y} p^y q^{n-y} I_{(0, 1, \ldots, n)}(y)$	$0 \le p \le 1$ n = 1, 2, 3, ... $(q = 1-p)$	Np	npq
Hipergeometrica	$f(y) = \frac{\binom{K}{y}\binom{M-K}{n-y}}{\binom{M}{n}} I_{(0, 1, \ldots, n)}(y)$	M = 1, 2, ... K = 0, 1, ..., M n = 1, 2, ..., M	$n\frac{K}{M}$	$n\frac{K}{M}\ \frac{M-K}{M}\ \frac{M-n}{M-1}$
Poisson	$f(y) = \frac{e^{-\lambda}\lambda^y}{y!} I_{(0, 1, \ldots)}(y)$	$\lambda > 0$	λ	λ
Geométrica	$f(y) = pq^y I_{(0, 1, \ldots)}(y)$	$0 < p \le 1$ $(q = 1-p)$	$\frac{q}{p}$	$\frac{q}{p^2}$
Binomial Negativa	$f(y) = \binom{r+y-1}{y} p^r q^y I_{(0, 1, \ldots)}(y)$	$0 < p \le 1$ R > 0 $(q = 1-p)$	$\frac{rq}{p}$	$\frac{rq}{p^2}$

Tabela 4. Distribuições contínuas de probabilidade

Distribuição	Função distribuição acumulada ou função densidade de probabilidade	Espaço Paramétrico	Média $\mu = E(Y)$	Variância $\sigma^2 = E[(Y-\mu)^2]$
Uniforme	$f(y) = \frac{1}{b-a} I_{(a,\ b)}(y)$	$-\infty<a<b<\infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp[-(y-\mu)^2/2\sigma^2]$	$-\infty<\mu<\infty$ $\sigma>0$	μ	σ^2
Exponencial	$f(y) = \lambda e^{-\lambda y} I_{(0,\ \infty)}(y)$	$\lambda>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gama	$f(y) = \frac{\lambda^r}{\Gamma(r)}\ y^{r-1}\ e^{-\lambda y}\ I_{(0,\ \infty)}(y)$	$\lambda>0$ $r>0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Beta	$f(y) = \frac{1}{B(a,b)}\ y^{a-1}(1-y)^{b-1}\ I_{(0,\ 1)}(y)$	$a>0$ $b>0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$
Cauchy	$f(y) = \frac{1}{\pi\beta\{1+[(y-\alpha)/\beta]^2\}}$	$-\infty<\alpha<\infty$ $\beta>0$	Não existe	Não existe
Lognormal	$f(y) = \frac{1}{y\sqrt{2\pi\sigma}}\ \exp[-(\log_e y-\mu)^2/2\sigma^2]\ I_{(0,\ \infty)}($	$-\infty<\mu<\infty$ $\sigma>0$	$\exp[\mu+(1/2)\sigma^2]$	$\exp\ [2\mu+2\sigma^2]$ $-\exp\ [2\mu+2\sigma^2]$
Exponencial Dupla	$f(y) = \frac{1}{2\beta}\ \exp\left(-\frac{\lvert y-\alpha\rvert}{\beta}\right)$	$-\infty<\alpha<\infty$ $\beta>0$	α	$2\beta^2$
Weibull	$f(y) = aby^{b-1}\ \exp[-ay^b]\ I_{(0,\ \infty)}(y)$	$a>0$ $b>0$	$a^{-1/b}\Gamma(1+b^{-1})$	$a^{-2/b}[\Gamma(2+b^{-1})$ $-\Gamma^2(1+b^{-1})]$
Logística	$f(y) = [1+e^{-(y-\alpha)/\beta}]^{-1}$	$-\infty<\alpha<\infty$ $\beta>0$	α	$\frac{\beta^2\pi^2}{3}$
Pareto	$f(y) = \frac{\theta X_0^{\theta}}{y^{\theta+1}} I_{(y_0,\ \infty)}(y)$	$y0>0$ $\theta>0$	$\frac{\theta_{y_0}}{\theta-1}$ $para\ \ \theta>1$	$\frac{\theta_{y_0^2}}{(\theta-1)^2(\theta-2)}$ $para\ \ \theta>2$
Gumbel ou Valor Extremo	$f(y) = \exp(-e^{-(y-\alpha)/\beta}$	$-\infty<\alpha<\infty$ $\beta>0$	$\alpha+\beta\gamma,$ $\gamma \approx .577216$	$\frac{\pi^2\beta^2}{6}$
Distribuição t	$f(y) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)}\ \frac{1}{\sqrt{k\pi}}\ \frac{1}{(1+y^2/k)^{(k+1)/2}}$	$k>0$	$\mu=0$ para $k>1$	$\frac{k}{k-2}$ $para\ \ k>2$
Distribuição F	$f(y) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\ \left(\frac{m}{n}\right)^{m/2}$ $y\frac{y^{(m-2)/2}}{[1+(m/n)y]^{(m+n)/2}}\ I_{(0,\ \infty)}(y)$	m, n + 1,2, ...	$\frac{n}{n-2}$ $para\ \ n>2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n}$ $N-4)$ $para\ \ n>4$
Distribuição Qui-Quadrado	$f(y) = \frac{1}{\Gamma(k/2)}\ \left(\frac{1}{2}\right)^{k/2}\ y^{k/2-1}\ e^{-(1/2)y}\ I_{(0,\ \infty)}$	k = 1, 2, …	k	2k

(b) Distribuição Normal Padrão

A variável Normal Padronizada é dada por $Z=\frac{y-\mu}{\sigma}$, a qual tem como função densidade de probabilidade:

$$f(z)=\frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}\ z^2\right\},\ z\in\Re$$

Média: 0

Variância: 1

A variável $Z=\frac{\bar{y}-\mu}{\sigma/(n)^{1/2}}$ também tem distribuição normal padrão, Z~N(0,1).

(c) Distribuição Qui-Quadrado (χ^2)

A variável $(n-1)\ \hat{\sigma}^2/\sigma^2$ tem distribuição qui-quadrado com ν=n-1 graus de liberdade. A estimativa $\hat{\sigma}^2$ advém de uma amostra aleatória da distribuição normal $y\sim N(\mu,\ \sigma^2)$. A função densidade de probabilidade desta distribuição é:

$$f(y)=\frac{\left(\frac{1}{2}\right)^{\upsilon/2}}{\Gamma(\nu/2)}\ y^{\nu/2-1}e^{-(1/2)y},\ \ y\geq 0\ \ e\ \ \nu\in N^*$$, em que:

Γ é a função matemática $\Gamma(m+1)=\int_0^\infty y^m e^{-y}dy$, com fórmula de recorrência $\Gamma(m+1)=m\Gamma(m)=m!$. Tem-se $\Gamma(1/2)=\pi^{1/2}$.

Média: ν

Variância: 2ν

(d) Distribuição t de Student

A variável $\bar{y}$ tem desvio padrão igual a $\sigma/(n)^{1/2}$. A variável $t = \frac{\bar{y} - \mu}{\hat{\sigma}/(n)^{1/2}}$ tem distribuição t de Student com n – 1 graus de liberdade, em que $\hat{\sigma}/(n)^{1/2}$ é uma estimativa de $\sigma/(n)^{1/2}$.

A distribuição t equivale a uma razão entre uma variável com distribuição normal padrão (Z~N(0,1)) e outra com distribuição qui-quadrado ($U \sim \chi^2_\nu$). Assim t = Z/(U/ν)$^{1/2}$, em que ν é o número de graus de liberdade.

A função densidade de probabilidade da distribuição t é:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)(n\pi)^{1/2}} \left(\frac{t^2}{n} + 1\right)^{-\left(\frac{t+1}{2}\right)}, \quad n > 0 \;\; e \;\; t \in \Re$$

Média: 0

Variância: $\frac{n}{n-2}$

Verifica-se que a distribuição t apresenta maior variância que a distribuição normal padrão, apresentando, portanto, cauda mais longa.

(d) Distribuição F de Snedecor

A variável aleatória $F = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2}$ tem distribuição F de Snedecor com n_1-1 e n_2-1 graus de liberdade, se $\sigma_1^2 = \sigma_2^2$. Têm-se diferentes distribuições $F_{(n1-1)(n2-1)}$ de acordo com o número de graus de liberdade. A distribuição F é, portanto, dada pela razão entre duas variáveis aleatórias independentes com distribuição qui-quadrado.

A função densidade de probabilidade da distribuição F e de outras distribuições contínuas relevantes são apresentadas na Tabela 4, conforme Mood et al. (1974). Em alguns casos, a parametrização é diferente daquela que foi apresentada nesse texto.

O ajuste de distribuição aos dados pode ser feito em vários *softwares* como o GenStat de Rothamsted – Inglaterra e o SAS da Carolina do Norte - EUA.

2.3 Teorema Central do Limite e Aproximação de Distribuições

O Teorema Central do Limite apregoa que se Y_1, Y_2, Y_3, ... Y_n são variáveis aleatórias independentes com médias $\mu_i \; e \; \sigma_i^2$, a distribuição de $\frac{S - \mu_s}{\sigma_s}$, sendo S = Y_1 + Y_2 + Y_3 + ... Y_n, é normal padrão (Z~N (0,1)) quando n $\rightarrow \infty$, ou seja, quando o tamanho da amostra é suficientemente grande.

Como a distribuição Binomial equivale à soma de variáveis aleatórias independentes Bernoulli, com n grande a média de S equivale a np e a variância a npq, de forma que tem-se $Z = \frac{Y - np}{(npq)^{1/2}}$, para Y com distribuição Binomial. Esta é a aproximação normal para a distribuição binomial. Esta aproximação é razoável quando p situa-se em torno de 0,5 e n é maior que 10, ou seja np $\geq$ 5. A aproximação normal da distribuição Poisson é adequada quando np $\geq$ 15.

Também a distribuição de Poisson pode ser considerada como uma distribuição limite da Binomial, quando p encontra-se próximo de zero e n tende ao infinito, com np mantendo-se constante. Esta aproximação é razoável quando n é grande e p < 0,1.

Devido às distribuições Binomial e Poisson serem discretas e a normal, contínua, recomenda-se a correção de continuidade de Yates. O Teorema Central do Limite garante também que quando médias de parcela de grande tamanho são utilizadas na análise de variância, a suposição de normalidade é aproximadamente atendida, mesmo que a distribuição dos dados não seja normal.

3 PRESSUPOSTOS DA ANÁLISE DE VARIÂNCIA E DE REGRESSÃO E SUAS VERIFICAÇÕES

No contexto do melhoramento genético, os efeitos de tratamentos (cultivares, variedades, linhagens, progênies, populações, híbridos, clones) podem ser considerados como fixos ou aleatórios, conforme detalhado no Capítulo 3. Em estatística, quando a variável resposta (y) é contínua (quantitativa) e os efeitos de tratamentos são considerados fixos, têm-se os modelos de análise de variância e de regressão, os quais são contemplados na maioria dos livros das várias áreas de aplicação da estatística: biometria, econometria, geoestatística e tecnometria (estatística aplicada à engenharia, física e química).

No caso da análise de variância (em que a variável indicadora (x) dos tratamentos é qualitativa), os modelos podem incluir efeitos fixos e aleatórios (por exemplo, tratamentos de efeitos fixos e blocos de efeitos aleatórios, no delineamento em blocos ao acaso), mas, as análises enfatizam os efeitos fixos, e os efeitos aleatórios são vistos como fatores de perturbação ou *nuisance*, não havendo interesse em testá-los. O efeito de tratamento não é uma variável casual, mas sim um tratamento introduzido sob controle, de forma que uma função do somatório do quadrado dos efeitos dos tratamentos não pode ser considerada como uma variância. O interesse principal da análise, no caso, é estimar as magnitudes dos efeitos de tratamentos e também das diferenças entre tratamentos.

Nos modelos de regressão, a variável indicadora dos tratamentos é quantitativa, ou seja, os tratamentos referem-se a quantidades, dosagens ou níveis de fatores quantitativos (neste caso, o uso da análise de variância é reprovável) e o interesse reside no estudo do comportamento da variável resposta em face do incremento das dosagens. Para as variáveis categóricas, têm-se os modelos de regressão logística e de análise de tabelas de contingência (Tabela 5).

Tabela 5. Classificação das variáveis resposta (y) e indicadora ou preditora (X) e seus modelos de análise associados

Variável y	Variável X	Modelo de análise
Contínua*	Categórica+	Análise de variância
Contínua*	Contínua*	Regressão
Categórica+	Contínua*	Regressão logística
Categórica+	Categórica+	Análise de tabelas de contingência

* Contínua e Quantitativa; + Categórica (Discreta) e Qualitativa

A situação em que os tratamentos são assumidos como efeitos aleatórios caracteriza os modelos de componentes de variância. Não há interesse na magnitude dos efeitos de tratamentos e seus contrastes, mas sim na magnitude da variância entre tratamentos (σ_t^2) e seu valor em relação à variação residual (σ^2), em geral, expressa por $\sigma_t^2/(\sigma_t^2+\sigma^2)$, fato que torna relevante a estimação de componentes de variância. Se os modelos incluem efeitos fixos, além do efeito aleatório de tratamento, estes são agora considerados de *nuisance.*

Assim, a rigor, estimar médias de tratamentos e realizar comparações de médias, quando os efeitos dos tratamentos são aleatórios não é teoricamente correto. Tanto na estimação dos efeitos fixos de tratamentos quanto na estimação de componentes de variância, os parâmetros de interesse referem-se a constantes (os parâmetros de locação e as variâncias paramétricas associadas às distribuições das variáveis aleatórias do modelo, respectivamente) que precisam ser estimadas. No caso em que os efeitos de tratamentos são aleatórios, eles podem ser preditos a partir dos dados observados e dos componentes de variância, caracterizando a predição de variáveis aleatórias.

Aos modelos de análise de variância com tratamentos de efeitos fixos, seguem-se os testes de médias ou testes de comparação múltipla. Os principais testes de comparações múltiplas são: t (ou LSD de Fisher ou forma F protegida de Fisher, quando aplicado somente quando o teste F é significativo), Tukey (ou HSD), Duncan, Scheffé, Dunnet, Newman-Keuls (ou Student-Newman-Keuls – SNK ou simplesmente Keuls), Bonferroni (ou LSDB) e Scott-Knott. Destes, o teste de Dunnet é adequado para comparação das médias dos tratamentos com as médias das testemunhas e os

demais permitem comparações entre as médias dos vários tratamentos. Dos menos rigorosos para os mais rigorosos, têm-se: t, Duncan, Newman-Keuls, Tukey, Scheffé. O teste de Scott-Knott é conservador ou rigoroso mas não tanto quanto os testes de Scheffé, Tukey e Newman-Keuls. De maneira geral, os testes de Newman-Keuls e de Scott-Knott devem ser preferidos. Maiores detalhes sobre comparações múltiplas podem ser encontrados em Steel e Torrie (1980), Ramalho et al. (2000), Pimentel Gomes (1987) e Resende (2002).

São quatro os pressupostos da análise de variância e de regressão: aditividade dos efeitos do modelo, normalidade dos erros, independência dos erros e homocedasticia ou homogeneidade de variância dos erros. Fica assim, explícita, a importância da análise de resíduos no contexto da análise de variância e de regressão. Erros ou resíduos referem-se a discrepâncias entre os valores observados e preditos pelo modelo.

3.1 Aditividade

A aditividade é um dos pressupostos mais importantes. Sua violação leva à inadequação do modelo aos dados ou dos dados ao modelo. A aditividade pode ser verificada pelo método de Tukey (1949), o qual desdobra a soma de quadrados do erro em um grau de liberdade para a não aditividade e (η - 1) graus de liberdade para o resíduo da não aditividade, onde η são os graus de liberdade do erro. O quadrado médio da não aditividade é testado contra o resíduo da não aditividade, pelo teste F.

A detecção de não aditividade pode revelar a necessidade de transformação de dados, sendo indicada a transformação logarítmica para os casos de modelos multiplicativos. Outra alternativa é o ajuste de modelos multiplicativos, conforme descrito no Capítulo 8.

3.2 Normalidade

A não normalidade dos dados e dos erros é restritiva apenas em situações muito drásticas, especialmente quando o teorema central do limite não se aplica, ou seja, quando a distribuição da média dos dados não converge para a distribuição normal, com o aumento do tamanho da amostra. Quando a distribuição dos erros é muito assimétrica (coeficiente de assimetria muito diferente de zero) e platicúrtica ou achatada (coeficiente de curtose muito negativo), a não normalidade afeta o nível nominal de significância dos testes estatísticos, de forma que o nível escolhido pelo pesquisador não é assegurado na prática, usualmente sendo maior que o nominal, podendo ser constatadas diferenças significativas entre tratamentos que, em realidade, não são diferentes.

De maneira genérica, o teste F de Snedecor é robusto a pequenos desvios da normalidade dos erros, fornecendo resultados satisfatórios quando se verificam distribuições aproximadamente normais. Na presença de grandes desvios da normalidade e de dados com estrutura não linear, a aplicação direta dos modelos lineares sobre os dados observados torna-se imprópria. Neste caso, há duas opções: (i) a transformação dos dados de forma que se tornem adequados (ou se moldem) ao modelo linear; (ii) a moldagem dos modelos aos dados. A abordagem (ii) é preferida e caracteriza técnicas pertencentes à classe dos modelos lineares generalizados (Mc Cullagh e Nelder, 1989; RESENDE, 2002). O uso destas técnicas conduz, na pior das hipóteses, aos mesmos resultados que a abordagem linear.

Variáveis dicotômicas e categóricas, por vezes, não apresentam normalidade de erros. Estas variáveis são, provavelmente, uma aproximação da variável real de interesse e, muitas vezes, as categorias surgem porque não é possível medir a variável real de interesse. Tomar as variáveis categóricas como normais é tanto mais apropriado quanto mais normais forem os escores. Assim, quanto maior o número de categorias, menor é a relevância da transformação das variáveis ou dos modelos para se adequarem às variáveis.

Existem vários procedimentos para avaliar a normalidade de dados. Dentre eles, destacam-se os testes não paramétricos de Kolmogorov-Smirnov, de Shapiro-Wilk e de aderência do qui-quadrado (χ^2), que são os mais utilizados. Os testes de assimetria e curtose também permitem inferir

sobre a normalidade. O teste de aderência do χ^2 permite verificar ajustamento de uma distribuição de freqüências de uma amostra a uma distribuição teórica, no caso a distribuição normal padrão. A estatística χ^2 é dada por $\chi^2 = \sum_{i=1}^{c} (O_i - E_i)^2 / E_i$, em que O_i são os valores observados na classe i da distribuição de freqüências, E_i refere-se ao valor esperado (sob H_o, ou seja, sob a hipótese de normalidade) na classe i, com base na distribuição normal padrão, e c refere-se ao número de classes. O valor calculado do χ^2 é comparado com o valor tabelado de uma distribuição χ^2 com (c – k – 1) graus de liberdade, adotando-se determinado nível de significância. K refere-se ao número de parâmetros estimados, no caso k = 2, tendo em vista que são estimados a média e o desvio padrão para o cômputo da variável padronizada z_i. Sendo o valor de χ^2 menor que o valor tabelado, aceita-se a hipótese de que os dados pertencem a uma distribuição normal.

O teste de Shapiro e Wilk (1965) pode ser aplicado segundo os seguintes passos, para o caso de amostras com tamanho $n < 50$:

a) ordenamento crescente das observações $x_1 < x_2 \ldots < x_n$.

b) cálculo da variância $\hat{\sigma}^2$ das observações.

c) obtenção de k: k = n/2 se n é par e k = (n-1)/2 se n é ímpar.

d) obtenção de:

$$b = a_n(x_n - x_1) + a_{n-1}(x_{n-1} - x_2) + ... + a_{n-k+1}(x_{n-k+1} - x_k),$$

em que os valores de a_{n-i+1} são tabelados para n = 3, 4 . . . 50.

e) obtenção da estatística de teste $W = b^2 / \hat{\sigma}^2$.

f) compara-se o valor calculado de W com o tabelado e sob a hipótese H_o de normalidade, aceita-se a normalidade, se W calculado $<$ W tabelado.

A função indicadora de W é ($0 < W \leq 1$), ou seja, varia no intervalo de zero a um, sendo que valores pequenos de W revelam que os dados não se adequam à distribuição normal, conduzindo à rejeição da hipótese H_0.

O teste de Komogorov-Smirnov, a exemplo do χ^2 de aderência, também compara as distribuições dos dados observados com a dos dados esperados sob suposição de normalidade padrão. Este teste baseia-se na maior diferença D entre as freqüências acumuladas observadas e esperadas e tende a ser mais poderoso que o teste χ^2, sobretudo no caso de pequenas amostras. A estatística do teste D é comparada com os valores críticos tabelados. Maiores detalhes sobre vários testes não paramétricos são apresentados por Siegel (1956).

Constatando-se grande desvio de normalidade e não se optando pelo uso das técnicas pertencentes à classe dos modelos lineares generalizados, resta realizar a transformação dos dados. Persistindo o problema, devem ser empregadas as provas de livre distribuição, ou seja, os procedimentos não paramétricos. Os testes não paramétricos devem ser usados como último recurso, visto que os testes paramétricos têm melhores propriedades estatísticas, principalmente o poder.

3.3 Independência

A suposição de independência dos erros deve ser obedecida no contexto da análise de variância. Isto porque os modelos de análise de variância assumem uma estrutura de covariância residual com covariâncias ou correlações seriais nulas.

As correlações seriais entre observações (ou erros) surgem a partir da tomada de dados seqüenciais no tempo e/ou no espaço em um mesmo tratamento ou em diferentes tratamentos. São, então, gerados dados temporalmente e/ou espacialmente dependentes. As conseqüências dessas correlações na análise de variância são sérias e podem conduzir a erros nas inferências sobre as médias.

Apesar de existir um teste formal (teste de Durbin-Watson) para avaliar a ocorrência de correlação serial entre os erros, é importante destacar que tal correlação resulta da não aleatoriedade na obtenção da amostra de dados. Assim, a suposição de independência pode ser verificada simplesmente observando se as unidades amostrais foram obtidas independentemente.

Neste caso, é necessário saber exatamente como foram obtidos os dados, tendo em vista o planejamento e execução da amostragem e o objetivo da investigação. A casualização dos tratamentos nas unidades experimentais pode conduzir à independência dos erros. O *software* Selegen-Reml/Blup permite avaliar a correlação serial dos resíduos em experimentos de campo por meio do modelo 113.

Em vários casos, são obtidos dados temporalmente e/ou espacialmente dependentes como nos estudos de dados longitudinais (medidas repetidas) e/ou dados de parcelas vizinhas em experimentos de campo. Neste caso, devem ser utilizados modelos espaciais que contemplem (por exemplo, via linhas e colunas auto-regressivas), em suas estruturas de variâncias, os erros correlacionados. Estes modelos são, via de regra, adotados nas análises de séries temporais e em geoestatística. Na análise de medidas repetidas, a presença de autocorrelação ou correlação serial (correlações aumentando com a diminuição da distância temporal entre amostras) pode indicar a necessidade do uso de modelos multivariados ou de outras estruturas de covariância, conforme descrito no Capítulo 9.

3.4 Homocedasticia

A homocedasticia ou homogeneidade de variância dos erros associados aos vários tratamentos é muito importante no contexto da comparação de médias, visto que os testes de comparação múltipla baseiam-se em diferenças mínimas significativas, dependentes de uma variação residual comum a todos os tratamentos. Assim, os erros contribuídos pelos vários tratamentos devem, todos, ser estimativas de uma variância populacional comum.

A violação de qualquer das outras suposições da análise de variância pode conduzir à heterogeneidade de variâncias dos erros. A violação da suposição de homogeneidade de variâncias é grave quando a distribuição dos erros é leptocúrtica (curtose positiva) ou quando existe assimetria, e, no caso de distribuição leptocúrtica, o teste F tende a não rejeitar a hipótese da nulidade (H_o), quando ela é falsa (Scheffé, 1959).

Existem vários testes para inferências sobre a existência ou não de homogeneidade de variâncias, como o de Bartlet, o de Hartley e o de Levene. O teste de Bartlet para comparação de variâncias é muito sensível à falta de normalidade dos erros, sendo por isto muito criticado. O teste de Hartley ou teste do F máximo tem como estatística de teste $F_{\max} = \hat{\sigma}^2_{\max} / \hat{\sigma}^2_{\min}$ com t e (b - 1) graus de liberdade, em que $\hat{\sigma}^2_{\max}$ e $\hat{\sigma}^2_{\min}$ referem-se à maior e à menor estimativa das variâncias residual (ou dentro de tratamentos) dos tratamentos ou amostras, t diz respeito ao número de tratamentos ou amostras e b, ao número de observações (ou blocos) por tratamento. O valor calculado do F_{max} é comparado com o valor tabelado da distribuição de F_{max} de Hartley. Como regra prática, tem sido aceito que quando F_{max} calculado é menor que 3, a heterogeneidade de variância não é problemática.

O teste de Levene consiste na análise de variância e teste F dos resíduos (valores observados menos valores preditos pelo modelo) dos vários tratamentos. Se o teste F não indicar diferenças significativas entre os resíduos dos tratamentos, aceita-se a homogeneidade de variância dos resíduos. Este teste tem grande aplicação (van Valen, 1978) e vem substituindo, vantajosamente, os demais. O *software* Selegen-Reml/Blup fornece os resíduos obtidos por vários modelos de análise no arquivo com extensão ".dev". Esses resíduos podem ser analisados pelo modelo 96, cuja análise de variância resultante fornece o teste de Levene.

No caso em que a heterogeneidade de variâncias é comprovada, têm-se as opções de padronização e de transformação dos dados. Segundo Steel e Torrie (1980), a heterogeneidade de variâncias pode ser classificada como regular e irregular. O tipo irregular caracteriza-se pela inexistência aparente de relação entre médias e variâncias. Por outro lado, a heterogeneidade regular surge de alguma forma de não normalidade, caracterizando-se por um relacionamento (correlação) entre médias e variâncias. Neste último caso, a transformação se aplica bem, desde que a distribuição dos dados seja conhecida.

O balanceamento do experimento, ou seja, a adoção de números iguais de repetições por tratamento é um grande passo na proteção contra os efeitos da heterogeneidade de variâncias (Scheffé, 1959).

No melhoramento de plantas, a heterogeneidade de variância residual entre tratamentos pode ocorrer. Neste caso, um procedimento BLUP com heterogeneidade de variância pode ser utilizado, conforme apresentado no Capítulo 3. O *software* Selegen-Reml/Blup informa sobre a heterogeneidade de variâncias residuais entre tratamentos por meio do arquivo com extensão ".het" e realiza a analise BLUP sob heterogeneidade via o procedimento BLUP-HET.

3.5 Transformações de Dados

As transformações de dados podem, em casos específicos, resolver os problemas de não normalidade, heterogeneidade de variâncias e não aditividade. Transformações apropriadas podem gerar dados com distribuição aproximadamente normal e com independência entre médias e variâncias, resultando em variâncias homogêneas. Os principais tipos de transformação são a logarítmica, a raiz quadrada e a arco-seno ou angular.

A transformação logarítmica estabiliza a variância, na situação em que as variâncias são proporcionais ao quadrado das médias dos tratamentos. Em alguns casos, pode contribuir para a normalização dos dados e para a adequação do modelo aditivo linear. Para uma variável Y com distribuição normal de média u e variância σ^2, o Log de Y tem variância aproximada de σ^2/u^2.

A transformação raiz quadrada é indicada para estabilizar ou homogeneizar a variância quando existe correlação entre média e variância e a variável refere-se a uma contagem, com distribuição de Poisson (esta distribuição tem média igual a variância). Neste caso, a variável transformada pode ser considerada com distribuição normal.

A transformação arco-seno é aplicável a dados com distribuição binomial, expressos em frações ou porcentagens, ocorrendo estabilização da variância. Em geral, quando todos os dados equivalem a porcentagens apenas na faixa de 30 a 70, a transformação provavelmente não seja necessária.

Para dados discretos, em geral, recomenda-se verificar a existência ou não de correlação entre as médias de cada tratamento e suas variâncias. Se for constatada tal correlação, deve-se identificar a distribuição (Binomial ou Poisson) dos dados e aplicar a transformação recomendada.

Uma maneira simples de decidir se uma transformação será efetiva consiste na verificação da proporção entre o maior e o menor valor do conjunto de dados. Se essa proporção for maior do que 20, a transformação será útil (Fernandez, 1992). Um procedimento mais formal para detecção da necessidade de transformação dos dados e para indicação da transformação adequada a ser usada, refere-se ao método da transformação potência de Box e Cox (1964). Essa transformação é dada por $Y_t = Y^{\gamma}$, se $\gamma \neq 0$ e $Y_t = \log Y$, se $\gamma = 0$. No caso, Y refere-se ao dado original e Y_t ao dado transformado. O parâmetro de transformação γ varia na amplitude de –2 a 2 e é determinado pela minimização da soma de quadrados residual.

A transformação potência pode ser implementada segundo os seguintes passos:

a) Estimar as médias de tratamentos (M) e seus desvios padrões (S);
b) Calcular os logaritmos de M e S;
c) Plotar log(S) contra o log(M) e verificar a linearidade dessa relação. Uma forte relação não linear indica que a transformação potência não será apropriada ao conjunto de dados e, nesse caso, indica-se a utilização de testes não paramétricos baseados nos ranks das observações;
d) Regressar log(S) em log(M) e testar a significância da relação linear. Se a regressão não for significativa (a 5 %), não há necessidade de transformação dos dados. Se a regressão for significativa, usar a estimativa do coeficiente de regressão (β) para obtenção do parâmetro γ;
e) Obter γ por meio de $\gamma = 1 - \beta$. Por exemplo, se $\beta = 2$, o valor de γ será –1 e, portanto, a transformação ideal será a recíproca (1/Y) dos dados.

Alguns valores de γ conduzem às transformações comumente usadas, conforme abaixo.

β	γ	**Transformação**
2,00	-1	Recíproca
1,00	0	Logarítmica
0,66	0,33	Raiz Cúbica
0,50	0,50	Raiz Quadrada

Quando a transformação potência é usada, perde-se um grau de liberdade no resíduo da análise, uma vez que o mesmo conjunto de dados foi usado para estimar o parâmetro γ e determinar a transformação apropriada. Os testes de significância e as comparações de médias devem ser realizados sobre os dados transformados. Mas as inferências e interpretações práticas devem ser realizadas na escala original, via transformação das médias e intervalos de confiança para a escala original. Em resumo, a determinação da transformação adequada, sem fazer a análise dos resíduos e/ou usar a transformação potência, nem sempre será efetiva.

4 MOMENTOS CENTRAIS DOS DADOS: USOS NA CARACTERIZAÇÃO DE DISTRIBUIÇÕES DE PROBABILIDADE E ANÁLISE GENÉTICA

Em estatística, os dados revelam mais informação quando são organizados e apresentados em tabelas de freqüência (distribuição de freqüências), histogramas e/ou gráficos de probabilidade normal. Uma tabela de freqüências é construída de forma a apresentar a fração dos dados que ocorre em cada intervalo ou classe de valores observados. Os valores desta tabela são utilizados para composição de histogramas. Quando os dados são ordenados em ordem crescente, tabelas de freqüência acumuladas ou distribuições de freqüência acumulada podem ser obtidas. O gráfico destas freqüências acumuladas constitui o gráfico de probabilidade normal.

Apesar do grande apelo visual, as mais importantes características dos histogramas e gráficos de freqüência acumulada (probabilidade normal) são expressas em algumas poucas estatísticas resumo: (i) as medidas de locação; (ii) as medidas de dispersão; (iii) as medidas de forma.

As medidas de locação informam onde as várias classes de dados se situam. Destas, a média, a mediana e a moda permitem inferir sobre o centro de distribuição dos dados. Os vários quantis informam sobre a locação das demais partes do conjunto de dados.

As principais medidas de dispersão são a variância, o desvio padrão e a amplitude ou intervalo interquartal. Tais medidas são descritivas da variabilidade dos dados. Dentre as medidas descritivas da forma de distribuição dos dados, as principais são o coeficiente de assimetria (*skewness*), de curtose (*kurtosis*) e o coeficiente de variação, o qual fornece informação acerca do comprimento da cauda de certas distribuições.

As principais medidas de posição, dispersão e forma de um conjunto de dados ou distribuição são dadas pelas estatísticas de primeira, segunda, terceira e quarta ordens. A estatística de primeira ordem, denominada esperança, centro de massa ou média (μ), é uma medida de posição associada ao primeiro momento dos dados. A estatística de segunda ordem refere-se à variância (σ^2) e é uma medida de dispersão associada ao segundo momento dos dados. Outras estatísticas que caracterizam a estrutura ou forma dos dados são aquelas de terceira e quarta ordens, denominadas assimetria (α_3) e curtose (α_4), respectivamente.

Formalmente, os momentos dos dados equivalem aos valores esperados de uma função de uma variável aleatória. Sendo Y uma variável aleatória e g(·) uma função com domínio e contradomínio reais, define-se expectância ou valor esperado g(·) da variável aleatória Y, a função E [g(Y)] dada por:

(i) $E[g(Y)] = \sum_{Y} g(Y) \cdot P_Y(y)$ se Y é uma variável aleatória discreta;

(ii) $E[g(Y)] = \int_{-\infty}^{\infty} g(Y)\ f_Y(y)\ dy$ se Y é uma variável aleatória contínua com função densidade de probabilidade $f_Y(y)$.

Assim, tem-se:

a) Se g (Y) = Y, então, E[g(Y)] = E(Y) = μ_Y: primeiro momento;

b) Se g (Y) = Y^2, então, E[g(Y)] = E(Y^2): segundo momento;

c) Se g (Y) = Y^3, então, E[g(Y)] = E(Y^3): terceiro momento;

d) Se g (Y) = Y^4, então, E[g(Y)] = E(Y^4): quarto momento;

e) Se g (Y) = (Y-0), então, E[g(Y)] = E(Y) = μ_Y: primeiro momento centrado em zero (média);

f) Se g (Y) = $(Y-\mu_Y)^2$, então, E[g(Y)] = $E(Y-\mu_Y)^2$=Var (Y): segundo momento centrado na média (variância).

Os momentos de uma variável aleatória ou de sua correspondente distribuição são as potências das esperanças. O r-ésimo momento de uma variável aleatória Y é usualmente indicado por M_r e definido por $M_r = E(Y^r)$ se a esperança existe. O r-ésimo momento central de uma variável aleatória Y em torno de a é definido como $E[(Y-a)^r]$. Se a = μ_Y, tem-se o r-ésimo momento central de Y em torno da média μ_Y. Assim:

$M_1 = E[(Y-\mu_Y)] = 0$: primeiro momento central;

$M_2 = E[(Y-\mu_Y)^2] = Var\ (Y)$: segundo momento central.

A variância de uma variável aleatória Y com esperança $E(Y) = \mu_Y$ é definida por:

(i) $\sigma_Y^2 = Var(Y) = \sum_Y (Y-\mu_Y)^2 P_Y(y)$ se Y é discreta;

(ii) $\sigma_Y^2 = Var(Y) = \int_{-\infty}^{\infty} (Y-\mu_Y)^2 f_Y(y)\ dy$ se Y é contínua.

Os parâmetros populacionais associados aos quatro momentos mencionados são dados por:

$\mu = \frac{1}{n} \sum y_i = \frac{1}{n}\left(\sum y_i - 0\right)$ = média, estatística de primeira ordem ou primeiro momento centrado em zero;

$\sigma^2 = \frac{1}{n} \sum (y_i - \mu)^2$ = variância populacional, estatística de segunda ordem ou segundo momento (centrado na média) dos dados;

$\alpha_3 = \frac{\frac{1}{n} \sum (y_i - \mu)^3}{\sigma^3} = \frac{m_3}{\sigma^3}$ = coeficiente de assimetria, associado ao terceiro momento (centrado na média) dos dados (m_3);

$\alpha_4 = \frac{\frac{1}{n}\sum (y_i - \mu)^4}{\sigma^4} = \frac{m_4}{\sigma^4}$ = coeficiente de curtose, associado ao quarto momento (centrado na média) dos dados (m_4).

Usando um conjunto de dados organizados em uma distribuição de freqüências, tem-se:

Dados (y)	**Freqüência (f)**	fy	$d = y - \hat{\mu}$	d^2	fd^2	fd^3	fd^4
3	20	60	-1,7	2,89	57,8	-98,26	167,04
4	30	120	0,7	0,49	14,7	-10,29	7,20
5	20	100	0,3	0,09	1,8	0,54	0,16
6	20	120	1,3	1,69	33,8	43,94	57,12
7	10	70	2,3	5,29	52,9	121,67	279,84
$\sum$	100	470	-	-	161,0	57,60	511,36

Usando estimadores expressos em termos de freqüências (dados agrupados em classes), têm-se os seguintes parâmetros populacionais:

$\hat{\mu} = \frac{\sum fy}{N} = \frac{470}{100} = 4{,}7$

$\hat{\sigma}^2 = \frac{\sum fd^2}{N} = \frac{161{,}0}{100} = 1{,}61$

$\hat{\sigma} = (1{,}61)^{1/2} = 1{,}27$: desvio padrão

$C.V. = \frac{\hat{\sigma}}{\hat{\mu}} = \frac{1{,}27}{4{,}7} = 0{,}2702$: coeficiente de variação

$m_3 = \frac{\sum fd^3}{N} = \frac{57{,}6}{100} = 0{,}576$

$\alpha_3 = \frac{m_3}{\sigma^3} = \frac{0{,}576}{(1{,}27)^3} = \frac{0{,}576}{2{,}045} = 0{,}28$

$m_4 = \frac{\sum fd^4}{N} = \frac{511{,}36}{100} = 5{,}1136$

$$\alpha_4 = \frac{m_4}{\sigma^4} = \frac{5{,}1136}{(1{,}27)^4} = \frac{5{,}1136}{2{,}59} = 1{,}97$$

Para a distribuição normal padrão, $\mu = 0;\ \sigma^2 = 1;\ \alpha_3 = 0;\ \alpha_4 = 3$. Neste caso em que a assimetria e a curtose são calculadas pelos momentos (centrados na média) em unidades de desvio padrão, valores de $\alpha_4 < 3$ indicam uma distribuição achatada denominada platicúrtica e valores de $\alpha_4 > 3$ indicam uma distribuição aguda (não achatada) denominada leptocúrtica.

No caso da distribuição normal não padronizada, tem-se tanto a assimetria quanto a curtose iguais a zero. Neste caso, valores positivos de curtose indicam curva leptocúrtica e valores negativos indicam curva platicúrtica.

Valores positivos de α_3 (assimetria) indicam uma distribuição assimétrica positiva ou concentrada a direita enquanto valores negativos indicam assimetria negativa ou concentrada a esquerda.

Para dados brutos (não agrupados em classes), os estimadores amostrais (amostra de tamanho N) de μ, σ, α_3 e α_4 são (Fisher, 1948):

Média

$$\hat{\mu} = \frac{\sum y_i}{N}$$

Variância

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\mu})^2}{N - 1}$$

Assimetria

$$\hat{\alpha}_3 = \frac{N \sum (y_i - \hat{\mu})^3}{(N-1)(N-2)\hat{\sigma}^3}$$

Curtose

$$\hat{\alpha}_4 = \frac{N(N+1)\sum (y_i - \hat{\mu})^4}{(N-1)(N-2)(N-3)\hat{\sigma}^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Outras importantes medidas de locação são dadas por:

Mediana: representa o ponto médio dos valores observados, quando os mesmos são arranjados em ordem crescente; dessa forma, metade dos dados encontra-se abaixo e metade acima da mediana.

Com dados ordenados na forma $y_1 \leq y_2 \leq \ldots y_n$, a mediana é calculada por meio de uma das equações:

$$M = \begin{cases} y_{\frac{n+1}{2}}, \ se\ n\ é\ impar, ou\ seja, o\ valor\ de\ y\ na\ posição\ (n+1)/2; \\ \left(y_{n/2} + y_{[(n/2)+1]}\right)/2, \ se\ n\ é\ par, ou\ seja, a\ média\ dos\ valores\ de\ y \\ nas\ posições\ (n/2)\ e\ (n/2+1). \end{cases}$$

A mediana pode ser facilmente visualizada a partir de um gráfico de probabilidade normal. Uma vez que o eixo y apresenta a freqüência acumulada, a mediana pode ser lida no eixo x como aquele valor associado à freqüência 50 % no eixo y.

A mediana é uma estatística robusta, pois é insensível a um alto valor errático (*outlier*). A média aritmética e a variância são sensíveis aos *outliers*.

Moda: é o valor mais freqüente ou comum no conjunto de dados.

Quartis: os quartis dividem o conjunto de dados em quartos. Com valores arranjados em ordem crescente, um quarto dos dados situa-se abaixo do primeiro quartil (Q_1) e um quarto dos dados situa-se acima do terceiro quartil (Q_3). Os quartis podem ser lidos facilmente a partir de um gráfico de probabilidade.

Quantis: referem-se à generalização da idéia da mediana e quartis para qualquer fração de dados, como em décimos (decis), dentre outras.

Média harmônica (m_H)

$$\frac{1}{m_H} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i} \Rightarrow m_H = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i}}$$

Média geométrica (m_G)

$$\log \ m_G = \frac{1}{n} \sum_{i=1}^{n} \log \ y_i$$

Uma importante medida de dispersão é o intervalo interquartal (IQR), o qual é dado pela diferença entre o terceiro e o primeiro quartis, ou seja, IQR = $Q_3 - Q_1$. De modo diferente da variância e desvio padrão, o intervalo interquartal não usa a média como centro da distribuição. Assim, é menos sensível aos *outliers* do que o é a variância.

O coeficiente de assimetria sofre maior influência dos *outliers* do que o sofrem a média e a variância, visto que está envolvido em sua fórmula, o cubo da diferença entre os valores absolutos e a média geral. Em geral, é usado apenas o sinal do coeficiente de assimetria e não a sua magnitude, para descrever a simetria da distribuição. Um coeficiente de assimetria positivo indica distribuição com cauda longa à direita, fazendo com que a mediana seja menor que a média. Se este coeficiente é próximo de zero, a distribuição é aproximadamente simétrica e a mediana situa-se próxima à média.

O coeficiente de variação (CV) é uma estatística alternativa ao coeficiente de assimetria, para descrever a forma da distribuição. Tal coeficiente pode fornecer indicativos de problemas com os dados. Um CV maior que 1 (ou 100 %) indica a presença de *outliers*, as quais podem impactar significativamente as estimativas finais.

O conjunto de estatísticas e técnicas gráficas relatadas permite inferir sobre a distribuição dos dados. Certas técnicas de estimação são mais adequadas quando a distribuição dos dados é normal ou Gaussiana. Para uma distribuição normal tem-se que a média, mediana e moda são coincidentes ou aproximadamente iguais e os valores de assimetria e curtose são nulos. Uma relação crescente entre moda, mediana e média (Moda < Mediana < Média) indica assimetria positiva.

Certos métodos estatísticos baseiam-se mais fortemente nas distribuições do que outros. Alguns métodos de estimação baseados em normalidade são ainda adequados quando os dados não apresentam distribuição normal.

A maioria das ferramentas empregadas em análise genética faz uso das estatísticas de primeira ordem (componentes de médias, como o BLUP dos valores genéticos, por exemplo) e segunda ordem (componentes de variância, obtidas via REML). Entretanto, as estatísticas de terceira (assimetria) e quarta (curtose) ordens podem também ser usadas na análise genética, especialmente na detecção e caracterização da natureza de interações gênicas, conforme demonstrado inicialmente por Fisher (1932).

A assimetria e a curtose podem ser usadas na análise e interpretação de diferenças genotípicas entre populações. Estas estatísticas são poderosas na revelação de interações do tipo multiplicativa. Para populações com distribuição normal, a assimetria e a curtose devem ser zero. A detecção de não normalidade é muito informativa e pode ser utilizada para inferências sobre o controle genético dos caracteres. Em ausência de interação gênica (epistasia), a assimetria deve ser zero. Assimetria significativamente maior ou menor que zero indica interação gênica complementar e duplicada (aditiva x aditiva), respectivamente. A curtose é sempre negativa (platicúrtica) ou próxima de zero na ausência de interação gênica e somente é positiva (leptocúrtica) na presença de interação gênica (Pooni et al., 1977; Choo e Reinbergs, 1982). A assimetria permite não apenas detectar a presença de interação gênica ou epistasia, mas, também, permite inferir sobre a natureza e causa desta interação. A estimação da assimetria, da curtose e seus desvios padrões são simples e devem ser usados em análise genética.

A análise genética tradicional em genética quantitativa baseia-se na normalidade (após eliminação de possíveis efeitos de escala) e assume um modelo genético infinitesimal, ou seja, grande número de locos não ligados e independentes governando o caráter. Na análise exploratória de dados, deve-se discernir bem entre *outliers* aleatórios e *outliers* genéticos, os quais são explicados pela epistasia. Ambos os tipos podem violar a normalidade dos dados. Outro fator genético que pode conduzir a desvios da normalidade é a segregação de genes de grande efeito.

Esse fator conduz tanto à assimetria quanto à curtose e o grau de curtose pode ser usado como um estimador do número de genes de grande efeito, embora esse método seja sensível à heterogeneidade de variâncias.

5 TÉCNICAS DE ANÁLISE EXPLORATÓRIA DE DADOS

Um bom profissional em análise de dados deve sempre examinar detalhadamente o conjunto de dados antes de produzir resultados estatísticos. As técnicas de análise exploratória de dados propiciam este exame, permitindo a identificação de padrões e características dos dados.

Dentre as características desejáveis da análise exploratória de dados, destacam-se: (i) maximização do conhecimento sobre um conjunto de dados; (ii) avaliação da estrutura dos dados; (iii) detecção de *outliers* ou observações discrepantes; (iv) verificação de relações entre variáveis; (v) inferências sobre distribuição dos dados.

A análise exploratória de dados pode ser dividida em quatro temas principais: (i) representações gráficas e uso de estatísticas descritivas do conjunto de dados; (ii) análise de resíduos e identificação de *outliers*; (iii) re-expressão dos dados via transformações ou padronização; (iv) uso de métodos robustos ou estatísticas resistentes.

O tema (i) envolve as representações gráficas em histogramas, gráficos de probabilidade normal, gráficos ramo e folhas, esquema *box-plot* e as estatísticas descritivas denominadas medidas de posição, de dispersão e de forma da distribuição. O *box-plot* (caixa de bigodes) pode ser usado na detecção de *outliers*, e o gráfico de ramo e folhas propicia filtrar as observações extremas. O tema (iii), transformação, foi abordado no tópico 3.5. A análise de resíduos (tema ii) será abordada em tópico específico, pois as técnicas que utilizam resíduos são determinantes para a análise exploratória.

Os métodos robustos de estimação (tema iv) fazem largo uso da mediana como uma medida de locação resistente e da dispersão quartal ou intervalo interquartal como uma medida de dispersão mais resistente. No contexto bivariado, o coeficiente de correlação entre ordens, ou correlação de

Spearman entre duas variáveis não é fortemente influenciado por pares extremos. Assim, é robusto em relação ao coeficiente de correlação linear de Pearson. Grande diferença de magnitude entre estes dois tipos de coeficiente de correlação pode revelar a presença de pares (de variáveis) extremos. Entretanto, uma alta correlação de Spearman não indica necessariamente que a relação entre duas variáveis é linear. A correlação de Spearman, entre duas variáveis, notadamente mais alta que a correlação de Pearson, pode indicar uma relação não linear entre tais variáveis. Como exemplo, duas variáveis X e Y, em que Y é dado por $Y = X^2$, apresentará um valor de correlação linear de Pearson próximo de 0, mas um valor de correlação de Spearman igual a 1.

A técnica da regressão robusta é complementar à técnica clássica baseada em quadrados mínimos. Fornece resposta similar à técnica clássica quando a relação entre as variáveis é linear e os erros apresentam distribuição normal. Porém, produz resultados significativamente diferentes em ausência de normalidade ou presença de *outliers* significativos.

6 ANÁLISE DE RESÍDUOS

Nenhuma análise estatística será perfeita sem a análise de resíduos. Vários tipos de discrepância entre os modelos assumidos e o conjunto de dados podem ser detectados via estudo de resíduos e seus componentes.

Os resíduos são definidos como desvios entre os valores observados e os preditos de acordo com um modelo assumido. Se as suposições acerca de um modelo são válidas, o gráfico dos resíduos versus valores preditos apresentará distribuição aleatória de pontos. Se o gráfico apresentar algum padrão sistemático inexplicável, então o modelo assumido não é apropriado. A análise de resíduos permite estudar a heterogeneidade de variâncias, a independência de erros e a presença de *outliers*.

6.1 Os Resíduos e o Ajustamento de Modelos

Um modelo teórico para análise de dados geralmente é da forma:

Dados = modelo + erros	
$y_{ij} = \mu + b_j + t_i + e_{ij}$	(delineamento de blocos ao acaso, nível de observações individuais)
$y = Xb + Zt + e$	(delineamento de blocos ao acaso, nível do vetor de observações y)

em que:

μ, b_j, t_i, e_{ij}: efeitos da média geral, de blocos, de tratamentos e de erros, respectivamente.

b, t, e: vetores de efeitos da média geral somada aos efeitos de blocos, dos efeitos de tratamento e de erros, respectivamente.

X, Z: matriz de incidência para os efeitos b e t, respectivamente.

A suposição acerca de uma determinada distribuição dos erros conduz à especificação de um modelo de forma ótima. Assumindo erros com distribuição normal, independentes (não correlacionados) e identicamente distribuídos (homocedasticidade ou σ_e^2 constante), obtêm-se estimadores não viciados e de variância mínima para os parâmetros de um modelo linear generalizado. Posteriormente, são estimados os parâmetros do modelo e/ou preditos os efeitos aleatórios, realizados testes de significância e de hipótese sobre os parâmetros do modelo. A análise dos resíduos permite verificar as suposições realizadas sobre os erros. Os próprios resíduos e sua variância são utilizados na obtenção da estimativa da variância do erro, a qual é usada na própria realização dos testes e nos procedimentos de predição.

Os resíduos são estimativas dos erros e são calculados com base no ajustamento do modelo. Neste caso, o modelo para descrever os dados é:

Dados = modelo ajustado + resíduo

$$y_{ij} = \hat{y}_{ij} + r_{ij} = \hat{y} + \hat{e}$$

$$y = X\hat{b} + Z\hat{t} + r$$

em que:

$\hat{b}$, $\hat{t}$, r : efeitos estimados e/ou preditos ($\hat{b}$, $\hat{t}$) e resíduo, respectivamente.

Os resíduos consideram, então, o procedimento de ajuste aplicado aos dados e são dados por:

$r_i = y_i - \hat{y}_i$	em termos individuais, ou
$r = y - X\hat{b} - Z\hat{t}$	em termos vetoriais.

O modelo ajustado descreve os dados de uma forma incompleta e a análise de resíduos pode contribuir para a melhoria do ajustamento por meio de: identificação da necessidade e uso de transformação dos dados; tratamento especial de valores discrepantes; inclusão de variáveis explicativas adicionais; diagnóstico de dependência espacial e uso de um modelo de análise espacial. O modelo ajustado pode ser melhorado até que esteja suficientemente próximo dos dados, de acordo com o objetivo do analista de dados. É desejável que a mesma técnica de ajuste, quando aplicada aos resíduos, produza um ajuste praticamente nulo, revelando que os resíduos têm um padrão ou comportamento completamente aleatório (Hoaglin et al., 1983).

O ajuste até aqui considerado contemplou a abordagem clássica da análise de dados. Entretanto, existem também as técnicas robustas e resistentes de análise, as quais necessitam de suposições mais fracas sobre os erros. Por exemplo, a presença de observações discrepantes pode exercer grande influência sobre o ajuste de modelos na abordagem clássica, mas não na robusta. Estes valores discrepantes referem-se aos *outliers* e também, especialmente na análise de regressão, aos pontos de alavancagem (“leverage”). Estes pontos influentes e de alavanca são dados atípicos muito úteis no processo de investigação.

Os resíduos são úteis na detecção de *outliers*, pois, no cálculo dos resíduos, faz-se a diferença entre valores observados e estimados, fato que revela a presença dos valores discrepantes. Assim, *outliers* nos dados e *outliers* nos resíduos são praticamente sinônimos. Com o uso de métodos resistentes, o ajuste não é significativamente afetado pelos *outliers*. Tais métodos contêm regras explícitas para o tratamento de *outliers*, as quais podem excluir o seu valor no ajustamento, sem, entretanto, excluir a sua presença. Como exemplo, *outliers* contribuem para a mediana por meio de sua presença e direção, mas não interferindo na sua grandeza (Hoaglin et al., 1983).

Os erros *e* são variáveis aleatórias não observáveis, e para verificar as hipóteses sobre a distribuição de erros, recorre-se à distribuição dos resíduos *r*. A distribuição dos erros e a distribuição dos resíduos quando se ajusta um modelo linear possuem semelhança assintótica, sendo que, em geral, a distribuição dos resíduos sob ajuste por quadrados mínimos está mais próxima da Normal do que o está a distribuição dos erros.

Quando os resíduos apresentam distribuição normal, mas com diferentes variâncias (por exemplo, mistura de duas distribuições normais com a mesma média e diferentes variâncias), a distribuição da mistura dessas distribuições tem cauda mais pesada que a normal. Neste caso, técnicas robustas de ajustamento são mais adequadas. Estas técnicas perdem pouca eficiência quando a distribuição dos erros não é normal e são ainda resistentes aos *outliers*.

6.2 Tipos de Resíduos e suas Variâncias

O resíduo $r = y - \hat{y} = y - X\hat{b} - Z\hat{t}$ tem variância:

$$\begin{aligned} Var(y - \hat{y}) &= Var(y) - 2Cov(y,\ \hat{y}) + Var(\hat{y}) \\ &= Var(y) - 2Var(\hat{y}) + Var(\hat{y}) \\ &= Var(y) - Var(\hat{y}) \end{aligned}$$

Sendo b um vetor de efeitos fixos e t um vetor de efeitos aleatórios (por exemplo, tratamentos genéticos como progênies ou linhagens não aparentadas), tem-se:

$$Var(y) = Z'\sigma_t^2 Z + \sigma_e^2$$
$$Var(\hat{y}) = Var(X\hat{b} + Z\hat{t}) = Z'\sigma_t^2 Z$$
$$Var(r) = Var(y) - Var(\hat{y}) = \sigma_e^2$$

A variância *Var(r)* é estimada por:

$Var(r) = \hat{\sigma}_e^2 = Var(y - X\hat{b} - Z\hat{t}) = [y'y - \hat{b}'X'y - \hat{t}'Z'y]/[N - p(X)]$, em que N é o número de dados e p(X) é o posto da matriz de incidência X. O estimador Var(r) propicia uma estimativa de σ_e^2.

São três os principais tipos de resíduos empregados na análise de resíduos:

(i) **Resíduos brutos ou não modificados**

$r_i = y_i - \hat{y}_i$ ou $r = y - \hat{y}$, os quais têm média zero e aproximada variância $Var(r) = \hat{\sigma}_e^2$.

(ii) **Resíduos padronizados**

$z_i = \dfrac{r_i}{[Var(r_i)]^{1/2}}$ ou $z = \dfrac{r}{[Var(r)]^{1/2}} = \dfrac{r}{\hat{\sigma}_e}$, os quais têm média zero e variância 1.

(iii) **Resíduos estudentizados ou Jacknife**

$z_i^* = \dfrac{r_i}{[Var(r_i^*)]^{1/2}}$ ou $z^* = \dfrac{r}{[Var(r^*)]^{1/2}} = \dfrac{r}{\hat{\sigma}_e^*}$.

A variância $Var(r_i^*)$ refere-se a uma variância $\sigma_{e(i)}^2$ para cada resíduo r_i, variância esta obtida com o ajuste do modelo sem considerar a observação y_i.

O resíduo padronizado é obtido através da transformação denominada padronização de uma variável aleatória, de forma que a mesma tenha esperança igual a zero e variância igual a 1. Esta transformação é obtida pela subtração da variável aleatória em relação a sua esperança (zero, no

presente caso) e divisão pelo seu desvio padrão (no caso, o parâmetro σ_e é substituído por seu estimador $\hat{\sigma}_e$). Sob modelos lineares adequados, z tem distribuição aproximada t de Student, sendo que a distribuição não é exata em função da dependência entre r_i e $\hat{\sigma}_e^2$, que participam no numerador e denominador de z_i, respectivamente. Para grandes tamanhos de amostra, z_i tem distribuição aproximadamente normal padrão.

Sendo aproximado por uma distribuição normal padrão, espera-se que os resíduos padronizados estejam em sua maioria (98 %) dentro do intervalo (-2,33; 2,33). Valores de resíduos fora desta faixa podem estar associados a observações discrepantes.

Em geral, não é necessário trabalhar com vários tipos de resíduos, sendo usual trabalhar com um ou dois tipos. Os resíduos padronizados e os estudentizados são semelhantes, sendo que os estudentizados são mais detalhadamente ajustados. Em algumas situações, os resíduos estudentizados são mais informativos que os padronizados, permitindo detectar a presença de pontos influentes. O programa SAS denomina z por resíduo estudentizado e z* por “Rstudent”.

6.3 Análise Gráfica de Resíduos

A representação gráfica dos resíduos é relevante na verificação dos padrões e distribuição dos resíduos bem como na detecção de valores discrepantes. O conhecimento dos resíduos permite a adoção de modelos de análise mais realísticos.

As representações gráficas dos resíduos pelas técnicas de caule e folhas (ou ramo e folhas) e box-*plot* (caixa de bigodes) permitem verificar o comportamento global da distribuição dos resíduos, tal como a assimetria (normalidade), a presença de valores extremos e os agrupamentos dos resíduos. O box-plot é uma excelente ferramenta, pois permite visualizar a locação, a dispersão, a simetria, as barreiras de *outliers* e os *outliers,* independentemente da forma de distribuição e do conjunto de dados. É construído com base na mediana e nos quartis da coleção de dados, fato que o torna resistente aos *ouliers* (Hoaglin et al., 1983).

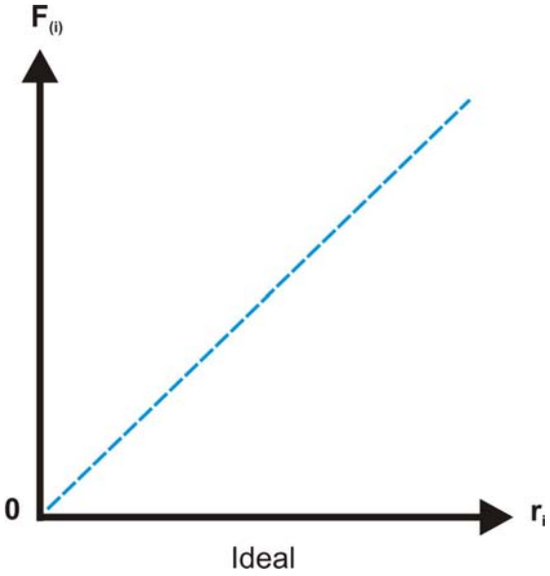
Outras representações gráficas relevantes são mostradas a seguir, as quais podem ser realizadas empregando-se quaisquer dos tipos de resíduo.

1) Gráfico de Probabilidade Normal

O gráfico de probabilidade normal é construído pela representação do i-ésimo resíduo (r_i) ordenado (em ordem crescente) no eixo X e o correspondente quantil da distribuição normal acumulada empírica (probabilidade $F_{(i)}$), na ordenada. Para que não haja evidência contra a normalidade dos resíduos, deve-se ter, no gráfico, uma reta aproximada. Em um modelo de efeitos fixos, a normalidade dos resíduos implica na normalidade dos próprios dados.

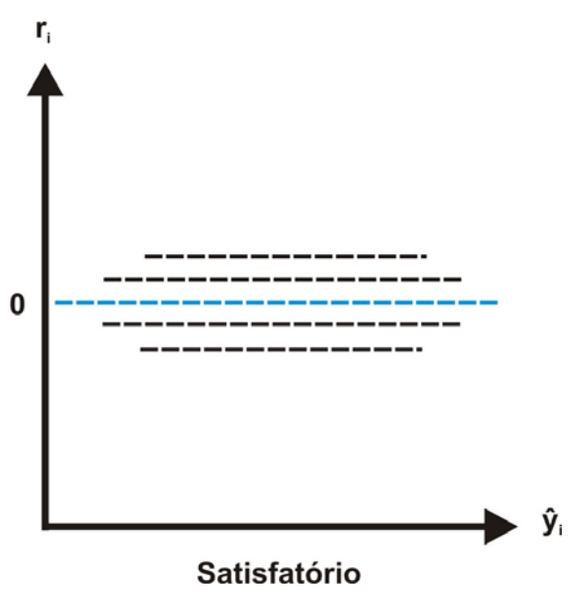
Existem outros métodos para verificação da normalidade como o estudo dos momentos e estatísticas descritivas (média, moda, mediana, assimetria, curtose) bem como a aplicação de testes como o de Kolmogorov, o de Shapiro-Wilk, dentre outros. Tais métodos foram descritos em tópicos anteriores.

O gráfico de probabilidade normal é também útil para revelar observações discrepantes em posições extremas e anômalas no gráfico. Pode revelar também a presença de resíduos sucessivos correlacionados, produzindo ondas no gráfico.

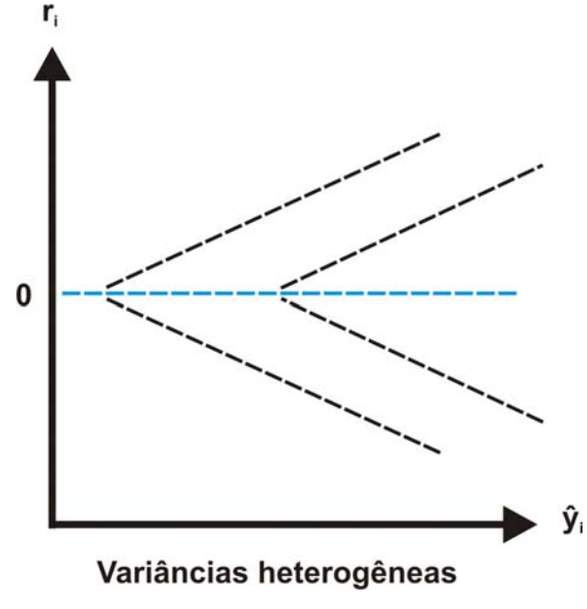


2) Gráfico dos Resíduos (r_i, z_i ou z_i^*) versus valores preditos $\hat{y}_i$

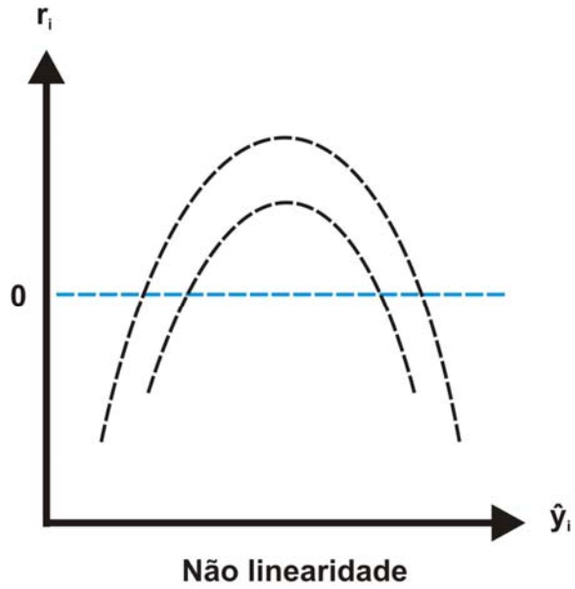
Neste gráfico, a ordenada refere-se aos resíduos e na abscissa encontram-se os valores preditos pelo modelo. Este gráfico é especialmente indicado para verificação da homocedasticidade. Em presença de homogeneidade de variância, este gráfico deve apresentar uma distribuição aleatória de pontos (linhas tracejada escuras no gráfico) em torno da média (linha tracejada azul no gráfico), sem qualquer padrão sistemático. Padrões sistemáticos neste gráfico podem indicar a presença de variâncias heterogêneas ou não linearidade. Este gráfico permite também detectar a presença de *outliers* (resíduos com valores absolutos maiores que 2,33).



Satisfatório



Variâncias heterogêneas

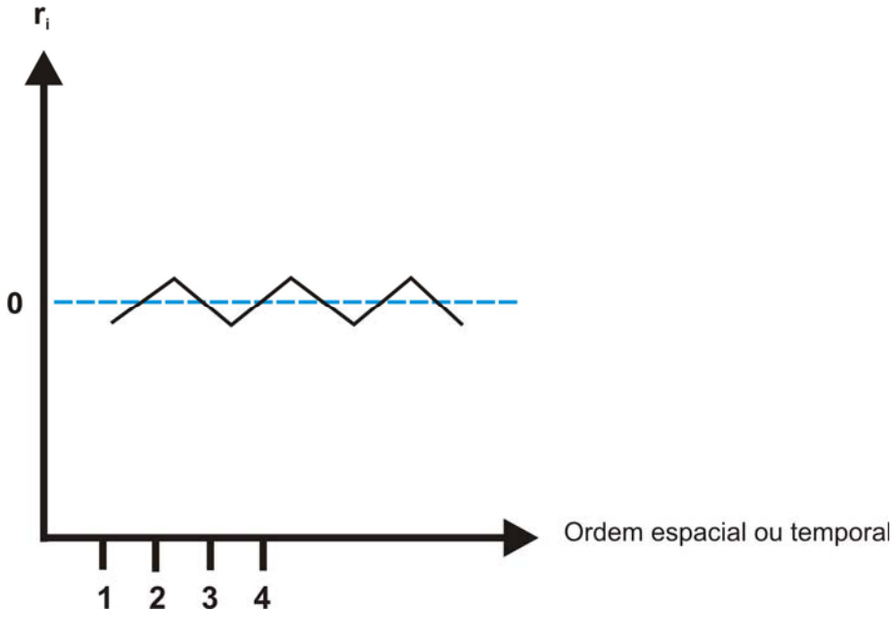


A presença de heterogeneidade de variâncias pode sugerir a necessidade de transformação dos dados. Neste caso, além das transformações citadas em tópico anterior, pode ser utilizado o procedimento generalizado de transformação, denominado transformação potência, dado por y em que é um parâmetro a ser determinado, geralmente considerado no intervalo de –2 a 2. Na presença de heterogeneidade de variâncias, estimadores de quadrados mínimos são ainda não viciados, mas não são de variância mínima.

A existência de uma relação entre tamanho do resíduo e valor predito revela que a variância dos resíduos é funcionalmente dependente da média. Este tipo de heterogeneidade de variância está usualmente associado a não-aditividade.

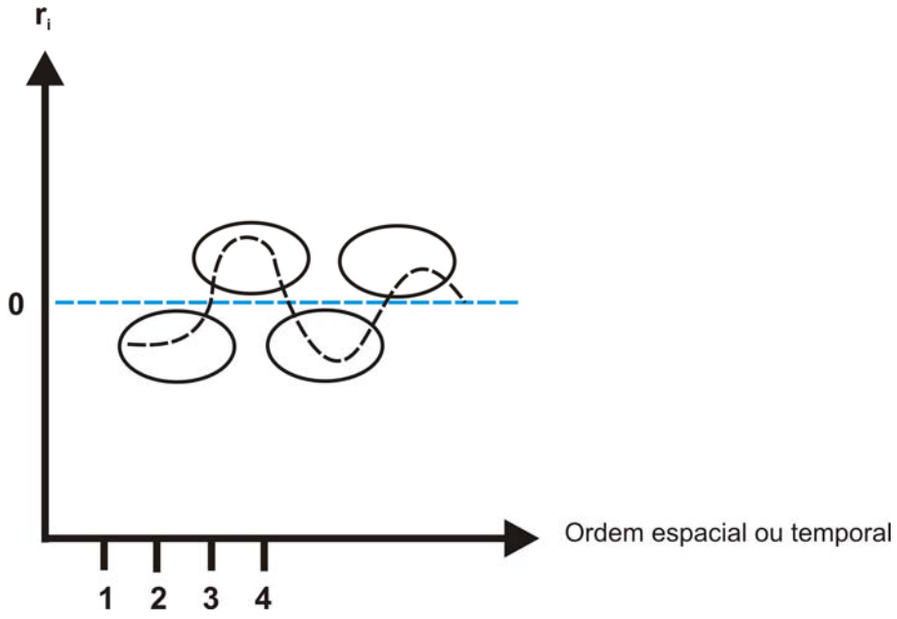
3) Gráfico dos Resíduos (r_i, z_i ou z_i^*) versus ordem espacial ou temporal das observações

Este gráfico tem a seguinte forma:



O gráfico apresentado, do tipo M, revela a presença de resíduos autocorrelacionados negativamente, ou seja, resíduos vizinhos com sinais trocados (mudança de sinal alternando rapidamente). Isto revela a falta de casualização na experimentação e/ou coleta de dados, ou, no caso de plantas, esta correlação negativa pode resultar de efeito de competição.

Em caso de autocorrelação positiva, o gráfico tem a forma:



Este gráfico revela resíduos de mesmo sinal ocorrendo em grupos. Tal autocorrelação positiva pode ser decorrente de dependência espacial ou temporal entre observações. A dependência espacial resulta de ambientes similares no solo onde estão plantas vizinhas.

6.4 Erros Correlacionados e Diagnósticos

Este tema é de fundamental importância na análise de resíduos pois, em presença de erros correlacionados, pode-se direcionar toda a metodologia de análise dos dados, em termos de modelagem, estimadores e abordagem estatística (estatística de variáveis aleatórias ou de variáveis mistas). Por exemplo, com dependência entre erros, a validade do teste F da análise de variância é seriamente afetada. Neste caso, no mínimo uma análise de variância com medidas repetidas ou uma análise de covariância com médias móveis deve ser realizada. Modelos mais complexos empregados no contexto da análise espacial são eficientes nesta situação.

Vários testes estatísticos podem ser usados para detectar a presença de autocorrelação de resíduos, sendo o teste de Durbin-Watson o mais utilizado. As hipóteses consideradas por este teste são:

H_o: $\rho_r = 0$: não existe autocorrelação de resíduos.

H_a: Hipótese alternativa: $\rho_r > 0$ ou $\rho_r < 0$: existe autocorrelação positiva ou negativa entre resíduos, respectivamente.

Para a hipótese H_a: $\rho_r > 0$, a estatística de teste é:

$$d = \frac{\sum_{i=1}^{n} (r_i - r_{i-1})^2}{\sum_{I=1}^{n} r_i^2}$$, em que:

r_i: refere-se aos resíduos ordenados no tempo ou espaço (posição da observação).

A estatística d não possui distribuição simétrica e considera erros em posições adjacentes, ou seja, considera observações com vizinhança de primeira ordem. Tal estatística apresenta relação direta com a autocorrelação de primeira ordem (ρ_r), dada por $\rho_r = (2 - d)/2$. Assim, $d = 2\ (1 - \rho_r)$ e, em ausência de autocorrelação espacial, d tem valor esperado igual a 2. Valores de d superiores a 2 indicam autocorrelação negativa ao passo que valores de d inferiores a 2 indicam autocorrelação positiva. Em geral, a heterogeneidade espacial ambiental causa autocorrelações positivas, ao passo que os efeitos de competição entre plantas vizinhas causam autocorrelações negativas.

O valor calculado de d deve ser comparado com os valores teóricos da distribuição de d com (n –1) graus de liberdade, em que n é o número de dados. A regra de decisão acerca de d é dada por:

Se $d \leq d_L$: há evidências para a rejeição de H_o e, portanto, aceita-se a hipótese de que a autocorrelação é positiva.

Se $d \geq d_U$: não há evidências para a rejeição de H_o e, portanto, a autocorrelação é nula.

Se $d_L < d < d_U$: o teste não é conclusivo.

Os valores de d_L (d inferior) e d_U (d superior) são fornecidos em tabelas associadas a certos níveis de significância. Por exemplo, um valor calculado d = 1,74 é maior que d_U = 1,69, para α = 5 % de significância. Assim, como $d > d_U$ não há evidências para a rejeição de H_o e conclui-se pela ausência de autocorrelação de resíduos. Para α = 5 % e n = 100, os valores tabelados são d_L = 1,65 e d_U = 1,69 (Koutsoyiannis, 1973).

Para a hipótese H_a: $\rho_r < 0$, a regra de decisão é análoga ao caso de $\rho_r > 0$, porém usando (4 - d) em vez de d. A estatística de teste é:

Se $(4 - d) \leq d_L$: há evidências para a rejeição de H_o e, portanto, aceita-se a hipótese de que a autocorrelação é negativa.

Se $(4 - d) \geq d_U$: não há evidências para a rejeição de H_o e, portanto, a autocorrelação é nula.

Se $d_L < (4 - d) < d_U$: o teste não é conclusivo.

A obtenção dos valores de autocorrelação residual e da estatística de Durbin-Watson podem ser obtidas pelo modelo 113 do *software* Selegen-Reml/Blup.

CAPÍTULO 3
ESTIMAÇÃO E PREDIÇÃO EM MODELOS LINEARES MISTOS

Este capítulo trata da modelagem estatística adequada aos experimentos de campo, os quais, via de regra, envolvem a avaliação de diferentes materiais genéticos.

1 MODELOS ESTATÍSTICOS E SELEÇÃO GENÉTICA

Em termos mais rigorosos, a seleção é um problema puramente estatístico, visto que na prática seleciona-se uma fração de indivíduos segundo seus valores genéticos os quais seguem uma distribuição de probabilidade. Pearson (1903) abordou o tema seleção derivando as médias e variâncias condicionais para a distribuição normal multivaridada e Lush (1931) foi o primeiro cientista a utilizar preditores de valores genéticos baseados em médias condicionais. Cochran (1951) estendeu as propriedades ótimas destes preditores para quaisquer distribuições.

No início da década de 1960, Lehman (1961) relatou que dois tipos de seleção haviam sido estudados pelos estatísticos: o Modelo I e o Modelo II de seleção, análogos aos correspondentes modelos de análise de variância. No Modelo I, os candidatos à seleção são de efeitos fixos, implicando na escolha entre tratamentos, representados por uma amostra aleatória de observações tomadas independentemente em cada tratamento. Segundo Henderson (1973), nenhuma teoria unificada foi desenvolvida para este tipo de seleção. Dudewicz (1976) aborda a questão de procedimentos de ordenamento e seleção neste contexto. Yates (1934) desenvolveu o método de estimação de efeitos fixos por quadrados mínimos.

No Modelo II, a seleção envolve candidatos considerados como variáveis aleatórias não observáveis pertencentes a uma determinada população. O Modelo II sempre foi considerado no melhoramento genético, associado aos índices de seleção envolvendo informações de parentes, desde o trabalho de Lush (1931).

O terceiro tipo de seleção foi negligenciado por estatísticos e melhoristas até o início da década de 1970. Este Modelo III de seleção, denominado Modelo Misto de Seleção (em analogia ao modelo misto de análise de variância) foi apresentado formalmente por Henderson (1973). Neste caso, os candidatos à seleção são variáveis aleatórias não observáveis pertencentes a mais que uma população, e o mérito de cada candidato é a soma da média da população mais o valor predito da variável aleatória associada ao candidato. Neste caso, a seleção depende, também, de efeitos fixos desconhecidos.

O modelo misto de seleção foi apresentado por Henderson (1973), mas, foi concebido por volta de 1949 pelo próprio Henderson, então aluno da disciplina de Matemática e Estatística, ministrada por Mood. A idéia surgiu de um problema estatístico simples apresentado por Mood: “Dado um escore de inteligência (QI) igual a 130, qual é a estimativa de máxima verossimilhança do verdadeiro QI do indivíduo? Assuma que o QI verdadeiro e o teste de QI tem distribuição normal com média 100, a variância do erro do teste do QI é 25 e a variância do QI verdadeiro é 225”. Henderson imediatamente aplicou o preditor de valor genético ($\hat{g}$) usado por Lush, dado por:

$$\hat{g} = \mu + \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}\ (y - \mu) = 100 + \frac{225}{225 + 25}\ (130 - 100) = 127$$

Esta era a resposta correta, mas, Mood não aceitou o método. Em vez disto ele sugeriu a maximização da distribuição conjunta do verdadeiro QI e do teste de QI. Lush mostrou-se surpreso e não convencido quando Henderson disse a ele que seu preditor era de máxima verossimilhança. Posteriormente, a dúvida de Lush foi justificada, uma vez que a função maximizada não era realmente uma função de verossimilhança. Na versão posterior do livro de Mood e Graybill (1963), o problema foi completamente modificado. Naquela época (1949), Henderson usou a idéia de Mood para derivar o método geral de seleção sob modelo misto denominado melhor predição linear não viciada ou não tendenciosa (BLUP) o qual surgiu da maximização da função densidade conjunta de y (valores fenotípicos) e g (valores genéticos) (Henderson, 1973).

O BLUP foi formalmente e amplamente divulgado a partir da década de 1970 (Henderson, 1973, 1975, 1976; Thompson, 1976, 1977, 1979). Segundo Searle (1997a), uma outra sigla correta para o BLUP é BLUERR (*Best Linear Unbiased Estimator of Realized Random Variables*). No entanto, tal autor comenta que essa sigla jamais seria tão difundida quanto BLUP.

2 PRINCÍPIOS DA AVALIAÇÃO GENOTÍPICA

2.1 Abordagem Conceitual e Prática Adequada para Avaliação de Materiais Genéticos

A avaliação de materiais genéticos em experimentos de campo tem dois objetivos: (i) inferir sobre os valores genotípicos de tais materiais; (ii) ordenar os materiais genéticos com base em seus valores genotípicos. Torna-se claro que não há interesse em estimar as médias fenotípicas dos materiais genéticos nos experimentos e sim estimar suas médias genéticas (ou valores genotípicos), ou seja, suas médias futuras, quando forem plantados novamente em plantios comerciais. Quando plantados comercialmente, mesmo que seja no mesmo local ou região da experimentação, os efeitos de blocos, parcelas e efeitos ambientais aleatórios não se repetirão. Como tais efeitos estão embutidos em alguma proporção nas médias fenotípicas, isto prova que tais médias não são adequadas para inferência sobre os valores genotípicos dos materiais genéticos. Assim, em trabalhos científicos e em catálogos de avaliação genética, a apresentação de médias fenotípicas

não é desejável e recomedável. Pelo contrário, devem ser apresentados os valores genotípicos livres dos efeitos ambientais. Esses valores genotípicos são os verdadeiros valores de cultivo e uso (VCU). Atualmente, os testes de VCU têm sido requisitados pelo Ministério da Agricultura para fins legais de recomendação, registro e proteção de cultivares. É importante relatar que o conceito de VCU é exatamente o conceito de valores genotípicos e, portanto, o uso de médias fenotípicas para inferência em testes de VCU não é recomendável.

Na estimação ou predição dos valores genotípicos, o mais importante é a escolha do método de estimação/predição. Esse método deve propiciar a inferência mais precisa e realista possível. E isso deve ser avaliado segundo parâmetros estatísticos adequados. No contexto da avaliação genotípica, o parâmetro estatístico mais importante é a acurácia seletiva ($\hat{r}_{\hat{g}g}$). Esse parâmetro refere-se à correlação entre o valor genotípico verdadeiro do material genético e aquele estimado ou predito a partir das informações dos experimentos de campo. A acurácia é tanto mais alta quanto menores forem os desvios absolutos entre os valores genéticos paramétricos ou verdadeiros e os valores genéticos estimados ou preditos. Tais desvios podem ser avaliados pela estatística erro quadrático médio (EQM). O erro quadrático médio de predição equivale à distância Euclideana média entre os estimadores e os correspondentes parâmetros e é dado por: $EQM = (Vício)^2 + PEV$, em que PEV refere-se à variância do erro de predição. Assim, o EQM congrega simultaneamente os conceitos de vício e precisão, os quais estão implícitos no conceito de acurácia. Um estimador ou preditor acurado do valor genotípico apresenta vício nulo ou pequeno e alta precisão (baixa variância do erro de predição). Minimizar o erro quadrático médio significa maximizar a acurácia. Assim, o método ideal de estimação ou predição dos valores genotípicos é aquele que minimiza EQM. Verifica-se que tal método pode ser viciado em pequeno grau, pois o que importa é minimizar a soma $(Vício)^2 + PEV$.

Na classe dos estimadores/preditores não viciados, a precisão é dada pelo parâmetro variância do erro de predição (PEV). A estatística PEV é relacionada à acurácia por meio das equações:

$\hat{r}_{\hat{g}g} = (1 - PEV / \sigma_g^2)^{1/2}$;

$PEV = (1 - \hat{r}_{\hat{g}g}^2)\sigma_g^2$, em que σ_g^2 é a variação genotípica entre os materiais em avaliação.

Desse modo, quanto menor o valor de PEV, maior a acurácia, e quanto maior a acurácia, menor o valor de PEV; ou seja, maior a precisão. Assim, na classe dos estimadores/preditores não viciados, a estratégia de minimizar PEV conduz também à maximização da acurácia. Mas, de maneira geral (relaxando a necessidade de não vício), o que deve ser minimizado é o EQM.

A média fenotípica, média aritmética ou média estimada pelo método de quadrados mínimos não é um estimador de mínimo EQM quando se tem mais que dois tratamentos ou materiais genéticos em avaliação. O trabalho de Stein (1955), que constituiu um verdadeiro paradoxo na Estatística, demonstrou que a média aritmética é estimador não admissível, isto é, que existem estimadores que propiciam menor erro quadrático médio ou menor risco que a média aritmética, quando mais que duas médias necessitam ser estimadas. Neste contexto, James e Stein (1961) apresentaram um estimador melhorado para a média populacional, que é dado por $M^* = k\ (\overline{Y}_{i..} - \overline{Y}_{...}) + \overline{Y}_{...}$, em que *k* é um fator regressor (ou de *shrinkage*) da média amostral de determinado tratamento ($\overline{Y}_{i.}$) sobre a média geral ($\overline{Y}_{...}$).

Os métodos (viciados ou não) que minimizam o EQM conduzem a estimadores/preditores do tipo *shrinkage*. Genericamente, um estimador do tipo *shrinkage* tem a forma de um escalar (variando entre zero e um) multiplicado por um vetor de médias estimadas por quadrados mínimos ou por máxima verossimilhança. Ou seja, para o caso balanceado, esse tipo de estimador multiplica as médias fenotípicas por um fator que varia entre zero e um, dependendo da confiabilidade que se tem nas médias fenotípicas estimadas. Interessante é que tal procedimento é normalmente usado na prática, porém de maneira empírica, por parte de melhoristas, economistas, administradores. Geralmente, esses profissionais sabem que a média fenotípica dos experimentos não se repete nos plantios com fins comerciais, ou seja, que a média em plantios é sempre menor que a média experimental. Então multiplicam as médias fenotípicas por empíricos coeficientes de confiabilidade. Tais profissionais estão de fato usando estimadores do tipo *shrinkage*, porém de uma maneira não otimizada. De qualquer forma, confirmam a necessidade do uso desse tipo de estimador em detrimento das médias fenotípicas. Nos casos dos testes genéticos, o fator de confiabilidade ótimo é função da herdabilidade ao nível de médias através das repetições.

Estimadores do tipo *shrinkage* começaram a ser usados por Lush (1931) no contexto do melhoramento animal associado ao método da melhor predição linear (BLP) e, posteriormente,

foram também usados no método da melhor predição linear não viciada (BLUP) conforme Henderson (1973; 1975) e Thompson (1976; 1979). Esses métodos assumem os efeitos de materiais genéticos como aleatórios e o BLUP é, adicionalmente, um preditor não viciado. Entretanto, conforme Stein (1955), para mais que dois tratamentos, estimadores do tipo *shrinkage* são necessários, independentemente se os efeitos forem tomados como fixos ou aleatórios. O estimador melhorado de James e Stein (1961) não necessita de qualquer suposição referente a efeitos fixos ou aleatórios, ou sobre distribuições das médias a serem estimadas (Efron e Morris 1977). Requer apenas o relaxamento da suposição de não vício. Este estimador é viesado, mas tem menor erro quadrático médio que o estimador de quadrados mínimos, em determinada região do espaço paramétrico. No contexto da avaliação genotípica, é importante relatar que o vício propiciado pelo estimador de James-Stein é pequeno e só pode existir quando o número de tratamentos é baixo (inferior a dez). À medida que o número de tratamentos aumenta, o estimador viesado torna-se não viesado e, por isso, o estimador de James-Stein é denominado como “aproximadamente não viesado”. Conforme Schaeffer (1999), a princípio, somente estimadores não viesados eram usados pelos estatísticos. Os desenvolvimentos teóricos, porém, evidenciaram que tais estimadores podem gerar estimativas fora do espaço paramétrico admissível. Assim, atualmente, procedimentos aproximadamente não viesados, desde que eficientes (de mínimo erro quadrático médio), têm sido considerados como os ideais.

Em síntese, na avaliação genotípica, a relevância maior não está na escolha entre efeitos fixos ou aleatórios dos efeitos de tratamentos, mas sim, na escolha de estimadores/preditores mais acurados e de mínimo erro quadrático médio, isto é, aqueles que propiciam as inferências mais corretas. Nesse sentido, ganham importância fundamental os estimadores de James e Stein (1961), também reconhecidos como estimadores do tipo *shrinkage*, os quais não assumem aleatoriedade, mas propiciam, com o aumento do número de tratamentos em avaliação, uma transição natural de efeitos puramente fixos para efeitos completamente aleatórios. E isso só depende do tamanho da amostra (número de tratamentos). Com grande número de tratamentos (diga-se > que 8), os estimadores de James-Stein e o método BLUP se equivalem (Resende e Duarte, 2007). Nesse caso, a metodologia BLUP é a melhor escolha pela facilidade de implementação e por poder ser estendida para o caso não balanceado.

O procedimento de estimação bayesiana, derivado em 1763 e, portanto, bem mais antigo do que o método de Stein, também minimiza o erro quadrático esperado. Por isso, o estimador de James-Stein é muito similar ao estimador de Bayes, tornando-se inclusive idênticos para grande número de tratamentos (Efron e Morris 1977). Por isso, são também denominados como estimadores de Bayes-Stein, Bayes empírico ou regra empírica de Bayes. Em inferência bayesiana não existe qualquer distinção entre efeitos fixos ou aleatórios, e os parâmetros a serem estimados são considerados variáveis aleatórias (Gianola e Fernando 1986), que devem ser estimadas considerando as incertezas a elas associadas. No caso de inferências sobre médias populacionais de tratamentos, sob o enfoque bayesiano, Box e Tiao (1973) apresentam como regressor a quantidade (1 – 1/*F*), a qual equivale ao regressor de James-Stein quando o número de tratamentos é elevado e também ao regressor usado pelo método BLUP no caso balanceado. A quantidade (1 – 1/*F*) é o fator de confiabilidade e é a própria herdabilidade ao nível de médias de repetições ou, em um sentido mais genérico ou estatístico, refere-se ao coeficiente de determinação dos efeitos de tratamentos ou coeficiente de determinação genotípico. Denominado dessa forma, evita-se o forte senso populacional associado à palavra herdabilidade, que foi originalmente conceituado como um parâmetro de uma população infinita e com cruzamentos ao acaso. Os estimadores do tipo *shrinkage* dependem essencialmente da estatística coeficiente de determinação dos efeitos de tratamentos e não do parâmetro populacional herdabilidade, embora em alguns casos ambos sejam equivalentes.

Enfim, pode-se constatar que, mundialmente, os estimadores de James-Stein só não são amplamente utilizados porque houve a grande revolução bayesiana. Com isso, tanto em econometria quanto em melhoramento animal, por exemplo, a abordagem vem sendo implicitamente implementada e utilizada via abordagem bayesiana. O renascimento da inferência bayesiana, a partir da década de 1970, foi, em grande parte, devido ao trabalho de Stein. Posteriormente ao seu trabalho, percebeu-se uma relação entre os estimadores do tipo *shrinkage* de James-Stein e os estimadores do tipo Bayes empírico. Robbins (1964) demonstrou que é possível conseguir o mesmo risco mínimo de Bayes, sem a necessidade de conhecimento da distribuição *a priori*, desde que o número de médias a serem estimadas seja grande. Isso é exatamente o que caracteriza o método Bayes empírico. O método Bayesiano legítimo depende do conhecimento de distribuições *a priori* de alguns parâmetros.

Deve-se enfatizar, portanto, que a filosofia associada aos estimadores do tipo *shrinkage* (James-Stein, Bayes empírico, Bayes legítimo, metodologia de modelos mistos com tratamentos de efeitos aleatórios ou BLUP), embora geral, coaduna-se perfeitamente às aplicações em genética e melhoramento de plantas. Isso porque a forma de se eliminar os efeitos residuais de ambiente, embutidos nos dados fenotípicos, é por meio de *shrinkage*, isto é, pela multiplicação do valor fenotípico corrigido, por uma função da herdabilidade do caráter sob seleção. Isto permite evitar a suposição implícita (na abordagem tradicional de inferências sobre médias fenotípicas, em que se consideram os efeitos de tratamentos como fixos) e irreal de que a confiabilidade ou herdabilidade (coeficiente de determinação) é máxima ($h^2 = 1,0$). É relevante frisar que, mesmo sob pequeno número de tratamentos em teste, a herdabilidade não poderia ser assumida desse modo, a menos que o número de repetições seja muito grande.

Assim, fica evidente a necessidade da correção das médias fenotípicas por meio de um coeficiente de determinação dado, aproximadamente, por $(1 - 1/F)$ para o caso de dados balanceados, em que F é a estatística de Snedecor da análise de variância. Essa correção pode ser aplicada da seguinte maneira, de acordo com o número de tratamentos avaliados:

(i) para três tratamentos: usar como regressor da média do tratamento a quantidade $k = 1 - [(T-2)/(T)]/F^*$, conforme James e Stein (1961). Esse regressor é centrado em zero e não na média geral, e deve multiplicar diretamente a média do tratamento e não o seu desvio em relação à média geral. No caso, F^* é também centrado em zero e não na média geral. Para o caso de três tratamentos, o regressor equivale então a $1 - 0,33/F^*$, resultado que coincide com o estimador *shrinkage* apresentado por Vinod (1976), no contexto da análise de regressão linear múltipla.

(ii) para quatro tratamentos ou mais: usar a expressão de James e Stein (1961) e Efron e Morris (1975, 1977), em que o regressor centrado na média geral é dado por $k = 1 - [(T-3)/(T-1)]/F$ (valores fornecidos na Tabela 6, reportada por Resende e Duarte (2007). Esses regressores devem multiplicar diretamente os desvios das médias dos tratamentos em relação à média geral, assim como se faz no procedimento de melhor predição linear não tendenciosa (BLUP).

A transição de um modelo de efeitos fixos (T = 1 ou 2) para um modelo de efeitos aleatórios de genótipos, em função do aumento no número de tratamentos, pode ser observada na Tabela 6 (construída considerando-se o caso de um modelo com efeitos fixos de genótipos, porém relaxando a condição de não-vício, na estimação dos referidos efeitos). Verifica-se que a partir de nove ou dez tratamentos praticamente existe uma equivalência entre modelo fixo com estimador do tipo *shrinkage* e modelo aleatório. Com número de tratamentos (*T*) maior que cinco, constata-se que o estimador *shrinkage* tende mais ao modelo aleatório do que ao modelo fixo, pois a fração a ser dividida por *F* é igual a 0,6, valor este mais próximo à unidade (modelo aleatório) do que de zero (modelo fixo). Dessa forma, se houver necessidade de escolha entre modelo aleatório e fixo, deve-se adotar o modelo aleatório quando $T > 5$, e o modelo fixo quando $T < 5$. Esse resultado concorda com Efron e Morris (1977), ao relatarem que o uso do estimador tipo *shrinkage* reduz substancialmente o risco na inferência quando o número de tratamentos é superior a cinco. Com $T = 5$, pode-se adotar qualquer dos modelos, mas, preferencialmente, o modelo de efeitos aleatórios, que é sempre conservador em relação ao modelo de efeitos fixos. Assim, o uso generalizado do modelo de efeitos aleatórios ou, equivalentemente, o uso de $(1 - 1/F)$ como fator de *shrinkage* é uma opção segura para o melhorista e para o experimentador. Os testes tradicionais de comparação de médias só seriam aceitáveis para números de tratamentos iguais a dois, três ou quatro. A rigor, segundo o número de tratamentos a serem comparados, pode-se adotar:

$T = 2$: teste *t* de Student;

$T = 3$: estimador de James-Stein centrado em zero;

$4 \leq T \leq 8$: estimador de James-Stein centrado na média geral;

$T > 8$: BLUP.

No procedimento BLUP lida-se com efeitos genéticos de tratamentos assumindo-os como aleatórios e com efeitos de macro-ambiente, alternativamente, como fixos ou aleatórios. Em algumas situações, alguns desses efeitos ambientais são considerados como fixos e, portanto, obtém-se estimativas BLUE desses efeitos, as quais são empregadas para se obterem as predições BLUP dos efeitos aleatórios. Uma questão que surge é se o uso de estimadores viesados, em lugar dos estimadores BLUE, obtidos via quadrados mínimos generalizados, propiciaria um melhoramento do

procedimento BLUP. Gianola (1990) considerou exatamente isso como uma forma de aprimorar o procedimento REML/BLUP; e, Weigel *et al.* (1991) trataram com mais detalhe essa questão, concluindo, via simulação, que ocorre sim um ligeiro melhoramento do método.

O efeito *shrinkage* sobre os desvios de ambiente, algumas vezes tratados como efeitos fixos, segue a mesma tendência mostrada na Tabela 6 (para observar essa tendência basta ler as entradas relativas ao número de tratamentos como se fossem referentes ao número de blocos, por exemplo). Entretanto, do ponto de vista do melhorista e geneticista, o que é conservador inverte-se em relação ao que foi comentado para o caso dos efeitos genotípicos. A fração $(1 - 1/F)$ penaliza os efeitos ambientais, fato que não é conservador para o melhorista. Assim, em caso de dúvida na consideração dos efeitos ambientais como fixo ou aleatório, o conservador é tratá-los como fixos. Note-se na Tabela 6 que, se o número b de blocos (repetições no delineamento de blocos ao acaso) for menor ou igual a cinco, é preferível tratar esses efeitos como fixos. Com $b > 10$, pode-se optar por tratá-los como efeitos aleatórios. Com $6 \leq b \leq 10$, seria melhor usar os estimadores de James-Stein, mas, como isso é difícil de ser implementado no contexto computacional do método REML/BLUP, a abordagem conservadora recomendaria tratar os efeitos de blocos como fixo. Em resumo, parece adequado tratar esses efeitos (blocos completos) como fixos, quando o número deles for menor ou igual a dez, e como aleatório, quando este número for maior que dez.

Quando os blocos são completos, tratá-los como fixo ou aleatório conduz basicamente à mesma herdabilidade ajustada (Resende, 2004). Nesse caso, quando o objetivo é seleção de tratamentos genéticos avaliados em todos os blocos, praticamente não existe diferença entre considerar blocos como efeitos fixos ou aleatórios. Por outro lado, quando os blocos são incompletos e com pequeno número de tratamentos em cada um, esses devem ser considerados como aleatórios, como forma de permitir a recuperação de informação genética interblocos. Outro caso é quando se objetiva a seleção de indivíduos (componentes das progênies) que estão em diferentes blocos completos. Nesse caso, a propriedade de não vício pode ser relevante, especialmente quando existe uma correlação entre os efeitos de blocos e o nível genético dos indivíduos que nele se desenvolvem (caso em que os melhores indivíduos são alocados nos melhores blocos). Se presente, este fato demanda a consideração dos efeitos de blocos completos como fixos.

Tabela 6. Valores dos regressores dos desvios das médias fenotípicas em relação à média geral, em experimentos balanceados, visando à obtenção de estimativas mais precisas de valores genotípicos para diferentes números de tratamentos (Resende e Duarte, 2007).

Número de tratamentos	Regressor[1]	Número de tratamentos	Regressor
3	1 - 0,33/*F**	14	1 - 0,85/*F*
4	1 - 0,33/*F*	15	1 - 0,86/*F*
5	1 - 0,50/*F*	16	1 - 0,87/*F*
6	1 - 0,60/*F*	17	1 - 0,88/*F*
7	1 - 0,67/*F*	18	1 - 0,88/*F*
8	1 - 0,71/F	19	1 - 0,89/*F*
9	**1 - 0,75/*F***	20	1 - 0,89/*F*
10	1 - 0,78/*F*	**21**	**1 - 0,90/*F***
11	**1 - 0,80/*F***	38	1 – 0,95/*F*
12	1 - 0,82/*F*	135	1 – 0,99/*F*
13	1 - 0,83/*F*	400	1 – 1/*F*

[1]- *F**: *F* de Snedecor centrado em zero, sendo que esse regressor deve multiplicar diretamente a média fenotípica e não o desvio; *F*: *F* de Snedecor centrado na média geral

Em suma, sob o ponto de vista do mínimo erro quadrático médio de predição, os efeitos de tratamentos devem ser considerados como aleatórios quando o número de tratamentos for superior a quatro, a menos que se usem os estimadores do tipo *shrinkage* de James-Stein.

Nesse contexto, deve-se acrescentar que assumir os efeitos de genótipos como aleatórios não implica, necessariamente, a obrigação de estimar componentes de variância explicitamente. Para o caso de dados balanceados, a herdabilidade (genericamente um fator de regressão ou coeficiente de determinação) necessária pode ser obtida diretamente do valor de *F* da ANOVA, o qual é normalmente obtido mesmo quando os efeitos de genótipos são tratados como fixos. Nesse caso, o valor genotípico predito é dado por $\hat{g} = VCU = m + (1 - 1/F)(\overline{Y_{i.}} - m)$, em que $\overline{Y_{i.}}$ é a média amostral da cultivar *i* (*i* = 1, 2, ... *T* tratamentos) e *m* é a média geral. Esta abordagem usa implicitamente todos os fundamentos genéticos da avaliação genotípica e atende, simultaneamente, qualquer das três abordagens científicas alternativas para o problema da estimação/predição de efeitos de tratamentos: metodologia de modelos mistos (tratamentos como efeitos aleatórios),

estimadores *shrinkage* de James-Stein (sem a necessidade de assumir efeitos fixos ou aleatórios) e abordagem bayesiana. É importante relatar que, sob o enfoque bayesiano, os efeitos de tratamentos são sempre considerados aleatórios, de forma que estimadores que promovem *shrinkage* são sempre utilizados, mesmo com pequeno número de tratamentos (menor que dez).

A estimação de componentes de variância tem sido realizada no contexto da estatística experimental mesmo com um reduzido número de graus de liberdade associado ao efeito em questão. Mead (1997) relata que um número mínimo de 8 graus de liberdade é adequado para estimação de uma variância residual. Esse número coincide com o número mínimo de T = 9 tratamentos usado para definir os efeitos de tratamentos como aleatórios (Tabela 6) e, portanto, estimar o componente de variância entre tratamentos. Também Piepho et al. (2003, pág. 316) relatam que as estimativas de componentes de variância são imprecisas quando se tem menos que 5 a 10 níveis do fator em questão.

Com a aceitação dos efeitos de tratamentos como aleatórios e/ou uso de estimadores do tipo shrinkage, os testes de comparações múltiplas entre médias de tratamentos não são naturalmente recomendados. Isso porque esses testes são derivados sob a suposição de efeitos fixos de tratamentos (Steel e Torrie 1980) e, também, porque são aplicados e produzem inferências sobre médias fenotípicas e não sobre médias genéticas (que só podem ser obtidas por meio de um estimador do tipo *shrinkage*). Ademais, tais testes (exceto o t de Student em algumas situações) assumem homogeneidade de variância residual entre tratamentos, suposição essa que pode não ser realista, conforme demonstrado na seqüência. Assim, não há lugar para os testes tradicionais de comparação múltipla no melhoramento genético. Abrindo mão do rigor, podem ser adotados apenas quando o número de tratamentos é menor que cinco e quando houver homogeneidade de variância residual. Também, o teste de Tukey, muito usado na prática, apresenta poder muito baixo para detectar diferenças significativas quando o número de tratamentos é grande (> 5) (Perecin e Barbosa, 1988; Ramalho et al., 2000; Resende, 2002a). Isto corrobora que tal teste só deve ser aplicado quando o número de tratamentos for da ordem de cinco ou menos.

Do ponto de vista da estatística clássica ou freqüentista, a média amostral é o melhor estimador não viciado (BLUE) da média populacional. O estimador apresentado, que introduz um regressor da média amostral do tratamento em relação à média geral, no caso (1 – 1/*F*), é viciado

(Gianola e Fernando 1986) quando o número de tratamentos é muito pequeno. Entretanto, propicia menor erro quadrático médio do que os estimadores BLUE (Henderson 1984), sendo, assim, vantajosos em relação aos estimadores não viciados (Efron 1975).

Vários autores (Hill e Rosenberger 1985; Stroup e Mulitze 1991; Piepho 1994; Resende *et al.* 1993, 1996; Piepho 1998; Smith et al., 2001; Duarte 2000; Duarte e Vencovsky 2001; Resende 1999, 2002, 2004; Resende e Duarte, 2007) também enfatizam a necessidade de se considerar os materiais genéticos como de efeitos aleatórios, mesmo se esses materiais forem tidos como de efeitos fixos por abordagens tradicionais. Se os efeitos ambientais (blocos, locais) podem ser considerados fixos ou aleatórios dependendo da situação, o mesmo não se pode dizer acerca dos efeitos genéticos, que são aleatórios por natureza. Assim, pela abordagem aqui defendida, os modelos em melhoramento genético e avaliação de cultivares, quase sempre (exceto com número de tratamentos inferior a cinco), devem ser aleatórios ou mistos com genótipos aleatórios. Nesse caso, a interação de genótipos com efeitos ambientais será também um efeito aleatório, fato que permite produzir inferências para toda a população de ambientes.

Apresenta-se a seguir um trecho (penúltimo parágrafo do trabalho de Smith et al. 2001, publicado no *Australian and New Zealand Journal of Statistics*, cujo título é "*The analysis of crop variety evaluation data in Australia"*).

> One possible contentious issue is the choice between fixed and random effects in the analysis. We have considered this in great depth and lament the lack of direction in the statistical literature. The standard text-book notion of effects being random if they have been sampled from a population and fixed if attention is confined only to those effects in the model (see Searle, 19971, for example) is unhelpful and can lead to a circular argument. In our opinion the choice depends on the aim of the analysis. In terms of variety effects our aim is to predict future performance. This is best achieved by assuming the variety effects to be random. Initially, plant breeders and evaluators were sceptical about the use of BLUPs. They now accept the method because the predictions have been more realistic. It is no longer true that yield gains observed by the farmers are substantially lower than that those predicted. We do not wish to predict environmental effects.

Os valores genotípicos das cultivares são variáveis aleatórias não observáveis e, portanto, desconhecidas. Realizações futuras dessas variáveis aleatórias podem ser preditas tratando-se os efeitos genotípicos como aleatórios. Assim, a consideração dos efeitos de tratamentos (cultivares) como aleatórios (ou o uso de estimadores tipo *shrinkage*) é essencial para se fazer seleção genética

e inferir sobre o VCU propriamente dito. Caso contrário, a seleção é fenotípica e não genética. Isso porque a forma de se eliminar os efeitos residuais de ambiente, embutidos nos dados fenotípicos, é por meio de *shrinkage*, isto é, pela multiplicação do valor fenotípico corrigido, por uma função da herdabilidade do caráter sob seleção.

Para a estimação ou predição dos valores genotípicos, métodos estatísticos e técnicas de genética quantitativa são utilizados. A modelagem estatística de caracteres quantitativos complexos propicia importantes informações das influências ambientais sobre os fenótipos e torna possível predizer fenótipos ainda não observados. Tais fenótipos ainda não observados são aqueles que se desenvolverão nos plantios comerciais e que são preditos pelos efeitos ou valores genotípicos quantificados a partir da análise estatística dos fenótipos observados no experimento. A análise estatística sob modelos com efeitos fixos de genótipos propicia inferências sobre os fenótipos que já foram observados. E o interesse do melhorista e do produtor rural é sobre os fenótipos que ainda não foram observados. Essa é uma diferença conceitual importante. Somente a análise estatística sob modelos do tipo *shrinkage* (por exemplo, com efeitos aleatórios de genótipos) propicia inferências sobre os fenótipos que ainda não foram observados, ou seja, permite a estimação da realização de uma variável aleatória não observada.

A questão de assumir efeitos de genótipos como fixos ou aleatórios merece, ainda, um comentário adicional. O melhorista de plantas, em geral, assume facilmente que os efeitos de blocos são aleatórios, mesmo quando se têm apenas dois ou três deles previamente escolhidos, em uma área restrita da região de plantio. Por outro lado, o mesmo melhorista geralmente tem dificuldade em assumir que cinco, dez, vinte ou mesmo cinqüenta cultivares possam ser tratadas como de efeitos aleatórios, mesmo sabendo-se que, intrinsecamente, os efeitos genotípicos são variáveis aleatórias não observáveis. Isso constitui um contra-senso que precisa ser vencido, pois esses geneticistas sabem que a média fenotípica não tem herdabilidade igual a 1,0. Ademais, Piepho et al. (2003, pág. 316) relatam que, mesmo quando se tem um fator intrinsecamente aleatório, mas com menos que cinco níveis, é melhor tratar esse fator como de efeito fixo. Essa afirmação é condizente com os resultados apresentados na Tabela 6. Ademais, tais autores (pág. 313 e 314) tratam como fixos os efeitos de blocos completos.

É interessante reportar que alguns autores (Cornelius et al. 1996; Pacheco et al., 1997; Cornelius e Crossa, 1999) assumem explicitamente um modelo com efeitos fixos de genótipos, mas utilizam estimadores do tipo *shrinkage* em busca de maior eficiência dos métodos preditivos. Cornelius et al. (1996) e Cornelius e Crossa (1999) demonstraram que os estimadores do tipo *shrinkage* apresentaram excelente performance quando aplicados a modelos completamente fixos com nenhum efeito aleatório, exceto o erro. Estimadores do tipo *shrinkage* são essencialmente estimadores Bayes empírico e comumente os estimadores bayesianos apresentam excelente performance mesmo quando investigados sob uma perspectiva puramente freqüentista (Carlin e Louis, 1996). Os livros de Theil (1971) e Gruber (1998) enfatizam o melhoramento da eficiência preditiva por meio do uso dos estimadores do tipo *shrinkage*.

2.2 Inferência Sobre Valores Genéticos nas Diferentes Situações

A inferência sobre a média genotípica ou VCU dos materiais genéticos deve então seguir a seguinte recomendação.

a) Situação de dados balanceados e homogeneidade de variâncias

As inferências sobre as médias genotípicas dos materiais genéticos devem ser dadas por $\hat{g} = VCU = m + (1 - 1/F)(\bar{Y}_{i.} - m)$, com desvio padrão dado por $SEP = (PEV)^{1/2} = (1 - 1/F)^{1/2}\sigma_e/(b)^{1/2}$, em que *b* é o número de repetições. Assim, a inferência intervalar sobre o VCU é dada por: $m + (1 - 1/F)(\bar{Y}_{i.} - m) \pm t(1 - 1/F)^{1/2}\sigma_e/(b)^{1/2}$, em que *t* é o valor tabelado da distribuição de Student, com graus de liberdade associados à estimativa da variância do erro de predição. O componente σ_e^2 equivale à variância residual comum a todos os tratamentos. Verificando-se a sobreposição dos intervalos de confiança de dois materiais genéticos, pode-se inferir se estes diferem ou não significativamente entre si. A diferença mínima significativa entre dois genótipos pode ser dada por $LSD = DMS = 1.4142\ t(1 - 1/F)^{1/2}\sigma_e/(b)^{1/2} = 1.4142\ t\ SEP$. É

interessante notar a relação entre a acurácia ($r_{\hat{g}g}$) e o fator de *shrinkage*: $r^2_{\hat{g}g} = (1 - 1/F)$. Assim, o fator de *shrinkage* equivale ao quadrado da acurácia e as médias genotípicas podem ser dadas alternativamente por $\hat{g} = VCU = m + r^2_{\hat{g}g}(\overline{Y}_{i.} - m)$.

O uso dessa abordagem não depende, necessariamente, da suposição de que os efeitos de genótipos sejam aleatórios ou de se adotar a abordagem bayesiana. Depende, apenas, de se assumir que a herdabilidade da média fenotípica de um material genético não equivale a 1,0 e que esta média não se repetirá exatamente no plantio comercial. Para isso, é necessária uma correção, a qual pode ser dada pelo estimador de James-Stein, sem a necessidade de assumir aleatoriedade.

b) Situação de dados desbalanceados e homogeneidade de variâncias

Nesta situação, existe uma herdabilidade no sentido amplo (coeficiente de determinação) em nível de médias $(\hat{r}^2_{\hat{g}g})$ e uma acurácia para cada material genético. Dessa forma, a estimação dos valores genotípicos pode ser realizada mais facilmente por meio do procedimento BLUP ou via abordagem bayesiana.

c) Situação de heterogeneidade de variâncias

No caso de heterogeneidade de variâncias residuais entre tratamentos, os testes de comparação múltipla também não são recomendados, tanto para o caso balanceado, quanto para o caso desbalanceado. Tais testes assumem variâncias residuais aproximadamente iguais (homogêneas) para todos os tratamentos. Isso, na prática, raramente se verifica devido ao fato dos diferentes níveis de segregação genética dentro de cada tratamento e, ou, porque cada tratamento experimenta ambientes mais ou menos heterogêneos entre eles, simplesmente em razão da amostragem ambiental, particularmente sob pequeno número de repetições. Nesse caso, além de $\hat{r}^2_{\hat{g}g}$ e acurácias diferentes, cada cultivar terá uma herdabilidade individual específica (h^2_{ei}), função da

variação residual dentro de cada tratamento (σ^2_{ei}). Nessa situação, um procedimento BLUP com variâncias heterogêneas (BLUP-HET) é que deve ser usado. Tal procedimento já se encontra implementado no *software* Selegen-Reml/Blup (Resende 2002b). Para o caso balanceado, uma abordagem alternativa é seguir a recomendação apresentada no tópico a, porém, computando-se para cada cultivar i o multiplicador $\hat{r}^2_{\hat{g}gi} = \frac{(F-1)}{(F-1)+\sigma^2_{ei}/\sigma^2_e}$, em lugar de $\hat{r}^2_{\hat{g}g} = 1-1/F$. Essa idéia é similar ao que se faz no teste *t* (de Student) para a comparação de duas médias, sob variâncias desiguais, ou seja, considera-se a heterogeneidade de variâncias.

Uma análise da heterogeneidade de variâncias residuais entre tratamentos pode ser feita observando-se a Tabela 7, reportada por Resende e Duarte (2007). A maior variância (74,0) foi constatada para a cultivar 6, e a menor (32,7) foi verificada para a cultivar 3. A relação entre essas duas variâncias é de 2,26. Esse valor é menor que 3,0 e, nesses casos, geralmente assume-se, pelo teste F máximo (de Hartley), homogeneidade de variâncias, aplicando-se, na seqüência, um teste de comparações múltiplas. Entretanto, em termos de avaliação genotípica, essas diferenças nas variâncias residuais são consideráveis. Os coeficientes de herdabilidade em nível de parcelas individuais variaram de 0,25 a 0,43, fato que corresponde a uma diferença relativa de 72 %. Em outras palavras, a cultivar 6 estaria segregando mais geneticamente (ou apresenta menor grau de homozigose) do que a cultivar 3 e/ou experimentou ambientes muito mais variáveis que a cultivar 3. Assim, a média fenotípica da cultivar 3 apresenta confiabilidade bastante superior (75 %) à da média da cultivar 6 (57 %), conforme explicitado pelas herdabilidades em nível médias de repetições. Isso conduziu, conseqüentemente, a diferentes acurácias seletivas entre os genótipos e, também, a diferentes predições (BLUP-HET) para os valores genotípicos.

Partindo-se da premissa defendida nesse texto, a abordagem tradicional de comparações múltiplas teria conduzido a inferências menos fidedignas sobre os valores genotípicos. A diferença entre essas duas cultivares, de 12 unidades de produção pela abordagem tradicional (55,0 – 43,0), pouco ultrapassa oito unidades pelo método BLUP-HET (53,7 – 45,2), o que equivale a uma redução de 30 %, a qual é bastante considerável.

Tabela 7. Resultados da avaliação genotípica de cultivares de soja considerando a presença de heterogeneidade de variâncias residuais entre os tratamentos.

Cultivar	Média fenotípica	Variância residual	Herdabilidade individual	Herdabilidade de médias	Média genotípica BLUP	Média genotípica BLUP-HET	Acurácia BLUP-HET
1	59,00	40,67	0,376	0,707	56,52	56,95	0,841
2	57,00	66,67	0,269	0,595	55,23	54,98	0,771
3	43,00	32,67	0,429	0,750	46,20	45,25	0,866
4	47,00	38,00	0,392	0,721	48,78	48,40	0,849
5	51,00	72,00	0,254	0,576	51,36	51,42	0,759
6	55,00	74,00	0,249	0,570	53,94	53,71	0,755
Homoced.[1]	52,00	54,00	0,312	0,645	52,00	-	0,803

[1]- Valores estimados sob a suposição de homogeneidade das variâncias residuais de tratamentos

3 AVALIAÇÃO DA QUALIDADE EXPERIMENTAL

A qualidade da avaliação genotípica deve ser inferida preferencialmente com base na acurácia. Em experimentos balanceados pode ser usada também a estatística F de Snedecor, conforme a Tabela 8 apresentada por Resende e Duarte (2007). Tal estatística contempla, simultaneamente, o coeficiente de variação experimental (*CVe*), o número de repetições (*b*) e o coeficiente de variação genotípica (*CVg*). A expressão $F = 1 + bCVg^2/CVe^2$ mostra que o valor *F* contempla os três parâmetros mencionados. Embora usado tradicionalmente para avaliar a qualidade experimental, o coeficiente de variação experimental isoladamente não é adequado para isso. O mesmo pode ser dito em relação ao índice de variação ou coeficiente de precisão ($Cve/(b)^{1/2}$) o qual considera simultaneamente o *Cve* e *b*, mas não o *CVg*. Os três parâmetros são necessários, pois a acurácia depende deles simultaneamente, conforme pode ser observado pela expressão alternativa $\hat{r}_{\hat{g}g} = \left[\frac{1}{1+(CVe^2/CVg^2)/b}\right]^{1/2}$.

Tabela 8. Valores adequados da estatística F de Snedecor, para os efeitos de tratamentos (cultivares), visando atingir determinada acurácia, e as categorias de precisão requerida na avaliação genotípica. Fonte: (RESENDE e DUARTE, 2007).

Acurácia	Classes de precisão	Valor de *F*
0,99	Muito alta	50,2513
0,975	Muito alta	20,2532
0,95	Muito alta	10,2564
0,90	**Muito alta**	**5,2632**
0,85	Alta	3,6036
0,80	Alta	2,7778
0,75	Alta	2,2857
0,70	**Alta**	**1,9606**
0,65	Moderada	1,7316
0,60	Moderada	1,5625
0,55	Moderada	1,4337
0,50	Moderada	1,3333
0,40	Baixa	1,1905
0,30	Baixa	1,0989
0,20	Baixa	1,0417
0,10	Baixa	1,0101

Os valores da estatística *F* que devem ser obtidos visando à determinação de VCU com elevada acurácia são destacados na Tabela 8. A expressão matemática que relaciona os valores adequados de F à acurácia requerida é dada por: $F = 1/(1 - \hat{r}_{\hat{g}g}^2)$. Verifica-se que, para atingir uma acurácia de 90 %, deve-se obter um valor de *F* igual a 5,26. Assim, este deve ser um valor de referência nos experimentos para avaliação de VCU. Esse valor independe da espécie e do caráter avaliado e pode ser tido como um valor padrão para qualquer cultura.

Para o processo de seleção em programas de melhoramento, devem ser buscados valores de acurácia acima de 70 %. Isso equivale a valores de *F* aproximadamente maiores que 2,0 (Tabela 8). Logo, valores de *F* inferiores propiciam baixa acurácia seletiva.

Outra estatística comumente calculada no contexto da avaliação genotípica, conforme proposta de Vencovsky (1987), é o coeficiente de variação relativa (*CVr* = *CVg/CVe*). Fixando-se o número de repetições, a magnitude de *CVr* pode ser utilizada para inferir sobre a acurácia e a precisão na avaliação genotípica. Valores de acurácia obtidos para diferentes valores de *CVr*, quando se é fixado o número de repetições, são apresentados na Tabela 9, apresentada por Resende e Duarte (2007). Nesta tabela estão destacados em negrito os valores de *CVr* necessários para se obter acurácias seletivas em torno de 90 %, para diferentes números de repetições empregados na experimentação. Tais valores variam, por exemplo, de 1,50 para o caso de duas repetições a 0,70 para o caso de dez repetições. Já com trinta ou quarenta repetições é possível alcançar-se uma acurácia de 90 % no processo seletivo, mesmo quando a relação *Cvg/Cve* é inferior a 0,40. Logo, valores adequados ou não de *CVr* devem ser inferidos em conjunto com o número de repetições. Vencovsky (1987) reporta que valores de *CVr* em torno da unidade são adequados para a experimentação com milho. De fato, confirma-se na Tabela 9 que o valor 1,0 é adequado para propiciar inferências com acurácias e precisões altas a muito altas. Entretanto, com número de repetições superior a cinco, valores de *CVr* menores que a unidade também podem propiciar elevadas acurácia e precisão.

A informação dessa tabela também pode ser usada de forma reversa, isto é, a partir do valor de *CVr* verificado, é possível inferir sobre o número de repetições adequado para se conseguir determinada acurácia seletiva.

Tabela 9. Valores de acurácia para diversos coeficientes de variação relativa (CV*r*), sob diferentes números de repetições (*b*). Fonte: (Resende e Duarte, 2007).

CVr [1]	***b***											
	2	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***20***	***30***	***40***
0,10	0,14	0,17	0,20	0,22	0,24	0,26	0,27	0,29	0,30	0,41	0,48	0,53
0,20	0,27	0,33	0,37	0,41	0,44	0,47	0,49	0,51	0,53	0,67	0,74	0,78
0,25	0,33	0,40	0,45	0,49	0,52	0,55	0,58	0,60	0,62	0,75	0,81	0,85
0,30	0,39	0,46	0,51	0,56	0,59	0,62	0,65	0,67	0,69	0,80	0,85	**0,88**
0,40	0,49	0,57	0,62	0,67	0,70	0,73	0,75	0,77	0,78	0,87	**0,91**	0,93
0,50	0,58	0,65	0,71	0,75	0,77	0,80	0,82	0,83	0,85	**0,91**	0,94	0,95
0,60	0,65	0,72	0,77	0,80	0,83	0,85	0,86	0,87	0,88	0,94	0,96	0,97
0,70	0,70	0,77	0,81	0,84	0,86	0,88	0,89	**0,90**	**0,91**	0,95	0,97	0,98
0,75	0,73	0,79	0,83	0,86	0,88	0,89	**0,90**	0,91	0,92	0,96	0,97	0,98
0,80	0,75	0,81	0,85	0,87	0,89	**0,90**	0,91	0,92	0,93	0,96	0,97	0,98
0,90	0,79	0,84	0,87	0,90	**0,91**	0,92	0,93	0,94	0,94	0,97	0,98	0,98
1,00	0,82	0,87	**0,89**	**0,91**	0,93	0,94	0,94	0,95	0,95	0,98	0,98	0,99
1,25	0,87	**0,91**	0,93	0,94	0,95	0,96	0,96	0,97	0,97	0,98	0,99	0,99
1,50	**0,90**	0,93	0,95	0,96	0,96	0,97	0,97	0,98	0,98	0,99	0,99	0,99
1,75	0,93	0,95	0,96	0,97	0,97	0,98	0,98	0,98	0,98	0,99	0,99	1,00
2,00	0,94	0,96	0,97	0,98	0,98	0,98	0,98	0,99	0,99	0,99	1,00	1,00
2,25	0,95	0,97	0,98	0,98	0,98	0,99	0,99	0,99	0,99	1,00	1,00	1,00
2,50	0,96	0,97	0,98	0,98	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00
2,75	0,97	0,98	0,98	0,99	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00
3,00	0,97	0,98	0,99	0,99	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00
3,25	0,98	0,98	0,99	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00
3,50	0,98	0,99	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	1,00
3,75	0,98	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
4,00	0,98	0,99	0,99	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00

[1]- Os valores *CVr* destacados em negrito são aqueles cuja acurácia seletiva atinge cerca de 90 %, para os respectivos números de repetições.

Na Tabela 10 são apresentados valores alternativos dos coeficientes de variação genotípica e experimental, necessários para se conseguir acurácias em torno de 90 % (equivalente a uma determinação de 81 %) na avaliação genotípica. Verifica-se que com os números de repetições normalmente empregados, entre dois e quatro, provavelmente não se atingem valores de acurácia desejados. Isso porque, nessas condições (baixo número de repetições), a acurácia de 90 % só é conseguida para caracteres com alta herdabilidade em nível de parcelas individuais (50 % a 70 %); fato que é improvável de ser realidade, visto que os caracteres de interesse são, em geral, quantitativos e de baixa herdabilidade. Nessa mesma situação, as magnitudes dos coeficientes de variação experimental são inadequadas para informar sobre precisão, mesmo quando valores baixos como 10 % são perseguidos. Assim, para os caracteres de produção, com herdabilidades individuais inferiores a 0,40, ao menos seis repetições seriam necessárias. E, neste caso, diversos valores do coeficiente de variação experimental podem ser aceitáveis, desde que mantenham uma relação próxima a 0,82 com o *CVe* (última coluna da Tabela 10, obtida de Resende e Duarte, 2007).

Tabela 10. Valores alternativos dos coeficientes de variação genotípica (CV*g*) e experimental (CV*e*), expressos em porcentagem, necessários para se conseguir acurácias em torno de 90 % (determinação de 81 %) na avaliação genotípica, sob diferentes números de repetições (*b*) (h^2_i é a herdabilidade em nível de parcelas individuais, obtida em cada caso)[1]. Fonte: (Resende e Duarte, 2007).

***b* = 2**			***b* = 3**			***b* = 4**			***b* = 5**			***b* = 6**		
***CVr* =1,50**		h^2_i	***CVr* =1,20**		h^2_i	***CVr* =1,01**		h^2_i	***CVr* =0,90**		h^2_i	***CVr* =0,82**		h^2_i
CVg	***CVe***		***CVg***	***CVe***		***CVg***	***CVe***		***CVg***	***CVe***		***CVg***	***CVe***	
5,0	3,0	0,74	5,0	4,0	0,61	5,0	5,0	0,50	5,0	6,0	0,41	5,0	6,0	0,41
10,0	7,0	0,67	10,0	8,0	0,61	**10,0**	**10,0**	0,50	**10,0**	**11,0**	0,45	**10,0**	**12,0**	0,41
15,0	**10,0**	0,69	**15,0**	**13,0**	0,57	15,0	15,0	0,50	15,0	17,0	0,44	**15,0**	**18,0**	0,41
20,0	13,0	0,70	20,0	17,0	0,58	**20,0**	**20,0**	0,50	**20,0**	**22,0**	0,45	20,0	24,0	0,41
25,0	17,0	0,68	**25,0**	**21,0**	0,59	25,0	25,0	0,50	25,0	28,0	0,44	**25,0**	**30,0**	0,41
30,0	**20,0**	0,69	30,0	25,0	0,59	**30,0**	**30,0**	0,50	**30,0**	**33,0**	0,45	30,0	37,0	0,40
35,0	23,0	0,70	35,0	29,0	0,59	35,0	35,0	0,50	35,0	39,0	0,45	35,0	43,0	0,40
40,0	27,0	0,69	**40,0**	**33,0**	0,60	40,0	40,0	0,50	40,0	44,0	0,45	40,0	49,0	0,40
45,0	**30,0**	0,69	45,0	38,0	0,58	45,0	45,0	0,50	45,0	50,0	0,45	45,0	55,0	0,40
50,0	33,0	0,70	50,0	42,0	0,59	50,0	50,0	0,50	50,0	56,0	0,44	50,0	61,0	0,40
55,0	37,0	0,69	55,0	46,0	0,59	55,0	54,0	0,51	55,0	61,0	0,45	55,0	67,0	0,40
0,6	0,4	0,69	60,0	50,0	0,59	60,0	59,0	0,51	60,0	67,0	0,45	60,0	73,0	0,40
0,7	0,4	0,70	65,0	54,0	0,59	65,0	64,0	0,51	65,0	72,0	0,45	65,0	79,0	0,40
0,7	0,5	0,69	70,0	58,0	0,59	70,0	69,0	0,51	70,0	78,0	0,45	70,0	85,0	0,40
0,8	0,5	0,69	75,0	63,0	0,59	75,0	74,0	0,51	75,0	83,0	0,45	75,0	91,0	0,40
0,8	0,5	0,69	80,0	67,0	0,59	80,0	79,0	0,51	80,0	89,0	0,45	80,0	98,0	0,40
0,9	0,6	0,69	90,0	75,0	0,59	90,0	89,0	0,51	90,0	100,0	0,45	90,0	110,0	0,40
1,0	0,6	0,69	95,0	79,0	0,59	95,0	94,0	0,51	95,0	106,0	0,45	95,0	116,0	0,40
1,0	0,7	0,69	100,0	83,0	0,59	100,0	99,0	0,51	100,0	111,0	0,45	100,0	122,0	0,40

[1]- Os valores *CVg* e *CVe* destacados em negrito são aqueles aproximadamente associados a *CVe* iguais a 10 %, 20 % e 30 %, valores esses tradicionalmente usados como referência na prática.

Para espécies perenes, em que várias safras são colhidas na mesma unidade experimental, as herdabilidades a serem consideradas na Tabela 10 referem-se às herdabilidades em nível de médias das várias safras. Em plantas anuais e perenes, nos casos em que os experimentos são

repetidos em vários ambientes, o mesmo valor *F* de referência (5,26) deve ser usado, porém, este valor deve ser aquele computado da análise conjunta dos vários experimentos, isto é: $F = QM_{Tratamentos} / QM_{Interação\ "Tratamentos\ x\ Experimentos"}$. Em testes de VCU, os experimentos têm sido instalados com duas ou três repetições e em dois ou três locais. Nesse caso, a eficiência desses testes depende da magnitude da interação genótipos x ambientes. Na Tabela 11 são apresentadas algumas informações referentes à acurácia desses ensaios.

Tabela 11. Valores de acurácia (Acur) para diversos coeficientes de herdabilidade em nível de parcelas individuais (h^2) e de médias (h^2m), correlação genotípica através dos ambientes (r_g), coeficientes de determinação dos efeitos da interação genótipos x ambientes (c^2int) e número de ambientes (L) e de blocos por ambiente (B)

h^2	C^2int	r_g	L	B	h^2m	Acur	h^2	c^2int	r_g	L	B	h^2m	Acur
0.1	0.203	0.33	2	2	0.27	0.52	0.3	0.61	0.33	2	2	0.48	0.69
0.1	0.203	0.33	2	3	0.31	0.56	0.3	0.61	0.33	2	3	0.48	0.70
0.1	0.203	0.33	3	2	0.35	0.59	0.3	0.61	0.33	3	2	0.58	0.76
0.1	0.203	0.33	4	2	0.42	0.65	0.3	0.61	0.33	4	2	0.65	0.80
0.1	0.203	0.33	4	3	0.48	0.69	0.3	0.61	0.33	4	3	0.65	0.81
0.1	0.081	0.55	2	2	0.29	0.54	0.3	0.245	0.55	2	2	0.56	0.75
0.1	0.081	0.55	2	3	0.36	0.60	0.3	0.245	0.55	2	3	0.60	0.78
0.1	0.081	0.55	3	2	0.38	0.62	0.3	0.245	0.55	3	2	0.66	0.81
0.1	0.081	0.55	4	2	0.45	0.67	0.3	0.245	0.55	4	2	0.72	0.85
0.1	0.081	0.55	4	3	0.53	0.73	0.3	0.245	0.55	4	3	0.75	0.87
0.1	0.05	0.67	2	2	0.30	0.54	0.3	0.15	0.67	2	2	0.59	0.77
0.1	0.05	0.67	2	3	0.38	0.61	0.3	0.15	0.67	2	3	0.64	0.80
0.1	0.05	0.67	3	2	0.39	0.62	0.3	0.15	0.67	3	2	0.68	0.82
0.1	0.05	0.67	4	2	0.46	0.68	0.3	0.15	0.67	4	2	0.74	0.86
0.1	0.05	0.67	4	3	0.55	0.74	0.3	0.15	0.67	4	3	0.78	0.88

segue

continuação Tabela 11

h^2	c^2int	r_g	L	B	h^2m	Acur	h^2	c^2int	r_g	L	B	h^2m	Acur
0.1	0.025	0.80	2	2	0.30	0.55	0.3	0.075	0.80	2	2	0.61	0.78
0.1	0.025	0.80	2	3	0.39	0.62	0.3	0.075	0.80	2	3	0.68	0.82
0.1	0.025	0.80	3	2	0.39	0.63	0.3	0.075	0.80	3	2	0.70	0.84
0.1	0.025	0.80	4	2	0.46	0.68	0.3	0.075	0.80	4	2	0.76	0.87
0.1	0.025	0.80	4	3	0.56	0.75	0.3	0.075	0.80	4	3	0.81	0.90
0.1	0.011	0.90	2	2	0.31	0.55	0.3	0.033	0.90	2	2	0.62	0.79
0.1	0.011	0.90	2	3	0.39	0.63	0.3	0.033	0.90	2	3	0.70	0.84
0.1	0.011	0.90	3	2	0.40	0.63	0.3	0.033	0.90	3	2	0.71	0.84
0.1	0.011	0.90	4	2	0.47	0.68	0.3	0.033	0.90	4	2	0.77	0.88
0.1	0.011	0.90	4	3	0.57	0.75	0.3	0.033	0.90	4	3	0.82	0.91
0.1	0	1.00	2	2	0.31	0.55	0.3	0	1.00	2	2	0.63	0.79
0.1	0	1.00	3	2	0.40	0.63	0.3	0	1.00	3	2	0.72	0.85
0.1	0	1.00	2	3	0.40	0.63	0.3	0	1.00	2	3	0.72	0.85
0.1	0	1.00	4	2	0.47	0.69	0.3	0	1.00	4	2	0.77	0.88
0.1	0	1.00	4	3	0.57	0.76	0.3	0	1.00	4	3	0.84	0.91
0.2	0.406	0.33	2	2	0.40	0.63	0.4	0.81	0.33	2	3	0.52	0.72
0.2	0.406	0.33	2	3	0.43	0.65	0.4	0.81	0.33	2	2	0.53	0.73
0.2	0.406	0.33	3	2	0.50	0.71	0.4	0.81	0.33	3	2	0.63	0.79
0.2	0.406	0.33	4	2	0.57	0.76	0.4	0.81	0.33	4	3	0.68	0.83
0.2	0.406	0.33	4	3	0.60	0.77	0.4	0.81	0.33	4	2	0.69	0.83
0.2	0.164	0.55	2	2	0.45	0.67	0.4	0.327	0.55	2	2	0.63	0.80
0.2	0.164	0.55	2	3	0.52	0.72	0.4	0.327	0.55	2	3	0.66	0.81
0.2	0.164	0.55	3	2	0.55	0.74	0.4	0.327	0.55	3	2	0.72	0.85

segue

conclusão Tabela 11

h2	c2int	rg	L	B	h2m	Acur	h2	c2int	rg	L	B	h2m	Acur
0.2	0.164	0.55	4	2	0.62	0.79	0.4	0.327	0.55	4	2	0.78	0.88
0.2	0.164	0.55	4	3	0.68	0.82	0.4	0.327	0.55	4	3	0.79	0.89
0.2	0.1	0.67	2	2	0.47	0.69	0.4	0.2	0.67	2	2	0.67	0.82
0.2	0.1	0.67	2	3	0.55	0.74	0.4	0.2	0.67	2	3	0.71	0.84
0.2	0.1	0.67	3	2	0.57	0.76	0.4	0.2	0.67	3	2	0.75	0.87
0.2	0.1	0.67	4	2	0.64	0.80	0.4	0.2	0.67	4	2	0.80	0.89
0.2	0.1	0.67	4	3	0.71	0.84	0.4	0.2	0.67	4	3	0.83	0.91
0.2	0.05	0.80	2	2	0.48	0.70	0.4	0.1	0.80	2	2	0.70	0.83
0.2	0.05	0.80	2	3	0.57	0.76	0.4	0.1	0.80	2	3	0.75	0.87
0.2	0.05	0.80	3	2	0.59	0.77	0.4	0.1	0.80	3	2	0.77	0.88
0.2	0.05	0.80	4	2	0.65	0.81	0.4	0.1	0.80	4	2	0.82	0.91
0.2	0.05	0.80	4	3	0.73	0.85	0.4	0.1	0.80	4	3	0.86	0.93
0.2	0.022	0.90	2	2	0.49	0.70	0.4	0.044	0.90	2	2	0.71	0.84
0.2	0.022	0.90	2	3	0.59	0.77	0.4	0.044	0.90	2	3	0.78	0.88
0.2	0.022	0.90	3	2	0.59	0.77	0.4	0.044	0.90	3	2	0.79	0.89
0.2	0.022	0.90	4	2	0.66	0.81	0.4	0.044	0.90	4	2	0.83	0.91
0.2	0.022	0.90	4	3	0.74	0.86	0.4	0.044	0.90	4	3	0.87	0.94
0.2	0	1.00	2	2	0.50	0.71	0.4	0	1.00	2	2	0.73	0.85
0.2	0	1.00	3	2	0.60	0.77	0.4	0	1.00	3	2	0.80	0.89
0.2	0	1.00	2	3	0.60	0.77	0.4	0	1.00	2	3	0.80	0.89
0.2	0	1.00	4	2	0.67	0.82	0.4	0	1.00	4	2	0.84	0.92
0.2	0	1.00	4	3	0.75	0.87	0.4	0	1.00	4	3	0.89	0.94

Verifica-se que, para caracteres com herdabilidade de parcelas individuais igual a 10 %, mesmo com três repetições e avaliação em quatro ambientes (locais e/ou anos), a acurácia máxima obtida foi de 76 %, em ausência de interação genótipos x ambientes. Para caracteres com herdabilidade de parcelas individuais igual a 20 %, mesmo com três repetições e avaliação em quatro ambientes (locais e/ou anos), a acurácia máxima obtida foi de 87 %, em ausência de interação genótipos x ambientes. Para caracteres com herdabilidade de parcelas individuais igual a 30 %, a acurácia máxima obtida foi de 91 %, com três repetições e avaliação em quatro ambientes (locais e/ou anos) e em ausência de interação genótipos x ambientes, o que é pouco provável de ser realidade. Assim, para caracteres com herdabilidades menores ou iguais a 30 %, o uso de duas ou três repetições por experimento é insuficiente para propiciar 90 % de acurácia seletiva, mesmo quando os experimentos são repetidos em dois, três ou quatro ambientes. Para caracteres com herdabilidades iguais ou maiores que 40 %, o uso de duas ou três repetições por experimento é suficiente para propiciar 90 % de acurácia seletiva, quando os experimentos são repetidos em dois, três ou quatro ambientes, desde que o nível de interação genótipos x ambientes seja moderado (Tabela 11).

Considere-se, como exemplo, um experimento de avaliação da produção de seis cultivares, testadas no delineamento inteiramente ao acaso, com quatro repetições. Os resultados da análise da variância são apresentados na Tabela 12. Os resultados dos coeficientes de variação e de outras estatísticas de interesse, associados ao experimento, são:

$CVe = \frac{\sigma_e}{m} \times 100 = 14,13\ \%$ (coeficiente de variação experimental);

$CPe = \frac{\sigma_e}{(b)^{1/2} m} \times 100 = \frac{CVe}{(b)^{1/2}} = 7,07\ \%$ (coeficiente de precisão experimental);

$CVr = CVg / CVe = [(F - 1) / b]^{1/2} = 0,67$ (coeficiente de variação relativa);

$CVg = CVr \times CVe \times 100 = 9,52\ \%$ (coeficiente de variação genotípica);

$\hat{h}_i^2 = \frac{(CVr)^2}{1 + (CVr)^2} = 0,31$ (herdabilidade de cultivares, em nível de parcelas individuais);

$\hat{r}_{\hat{g}g} = (1 - 1/F)^{1/2} = 0,8030$ (acurácia da avaliação genotípica).

Com base nesses resultados, as seguintes inferências podem ser realizadas:

(i) A estatística *F* de Snedecor, equivalente a 2,81, propicia uma acurácia seletiva de cerca de 80 %. Esta acurácia, embora possa ser classificada como alta (Tabela 8), não atinge o mínimo adequado (90 %) para experimentos de VCU;

(ii) O coeficiente de variação relativa (*CVr*) equivaleu a 0,67. Para o caso de quatro repetições este coeficiente teria que atingir o valor 1,0 para propiciar uma acurácia de 90 % (Tabela 9). Portanto, para testes de VCU, o valor de *CVr* observado neste experimento foi baixo;

(iii) O coeficiente de variação experimental (*CVe*) equivaleu a 14,13 %. Embora recomendações tradicionais considerem este valor satisfatório, para testes de VCU com quatro repetições, o coeficiente obtido pode ser considerado alto. Conforme mostrado na Tabela 10, o valor de *CVe* para propiciar uma acurácia seletiva de 90 % deveria ser igual ao valor do *CVg*, isto é, 9,52%;

(iv) A herdabilidade de cultivares, em nível de parcelas individuais, equivaleu a 0,31. Neste caso, para se atingir uma acurácia seletiva de 90 % seria necessário usar mais de seis repetições (Tabela 10) ou, mais especificamente, entre nove ou dez repetições, haja vista o valor de *CVr* (Tabela 9).

Tabela 12. Análise da variância da produção de grãos de seis cultivares, em quatro repetições (Machado *et al.,* 2005)

Fontes de Variação	**Graus de Liberdade**	**Quadrados Médios**	***F* de Snedecor**
Cultivares	5	152,0000	2,8148
Resíduo	18	σ_e^2 = 54,0000	-

Média geral = m = 52,0

4 PROCEDIMENTO ÓTIMO DE AVALIAÇÃO GENOTÍPICA E SIGNIFICÂNCIA DOS EFEITOS DO MODELO

A avaliação genotípica compreende a estimação de componentes de variância (parâmetros genéticos) e a predição dos valores genotípicos. Além da utilidade no processo de predição dos valores genéticos, as estimativas dos parâmetros genéticos tais quais a herdabilidade e correlações genéticas são fundamentais para o delineamento de eficientes estratégias de melhoramento. A experimentação de campo, via de regra, está associada a desbalanceamento de dados devido a vários motivos tais quais perdas de plantas e parcelas, desiguais quantidades de sementes e mudas disponíveis por tratamento, rede experimental com diferentes números de repetições por experimento e diferentes delineamentos experimentais, não avaliação de todas as combinações genótipo-ambiente, dentre outros. Em função disso e do que foi exposto no tópico 2.1, o procedimento ótimo de avaliação genotípica refere-se ao REML/BLUP (máxima verossimilhança residual ou restrita/melhor predição linear não viciada), também denominado genericamente de metodologia de modelos mistos. Estes procedimentos lidam naturalmente com o desbalanceamento conduzindo a estimações e predições mais precisas de parâmetros genéticos e valores genéticos, respectivamente.

O procedimento ótimo de seleção é o BLUP para os efeitos genéticos aditivos (a), de dominância (d) e genotípicos (g), dependendo da situação. O BLUP é o procedimento que maximiza a acurácia seletiva e, portanto, é superior a qualquer outro índice de seleção combinada, exceto aquele que usa todos os efeitos aleatórios do modelo estatístico (índice multiefeitos, conforme Resende e Higa, 1994), o qual é o próprio BLUP para o caso de dados balanceados (Resende e Fernandes, 1999). O BLUP permite também o uso simultâneo de várias fontes de informação tais quais aquelas advindas de vários experimentos instalados em um ou vários locais e avaliados em uma ou várias colheitas. O BLUP individual utiliza todos os efeitos do modelo estatístico, contempla o desbalanceamento, utiliza o parentesco genético entre os indivíduos em avaliação e, considera a coincidência entre unidade de seleção e unidade de recombinação. A predição usando BLUP assume que os componentes de variância são conhecidos. Entretanto, na prática, são necessárias estimativas fidedignas dos componentes de variância (parâmetros genéticos) de forma a se obter o

que se denomina BLUP empírico (Harville e Carriquiry, 1992). O procedimento recomendado para estimação de componentes de variância é o da máxima verossimilhança restrita (REML), desenvolvido por Patterson e Thompson (1971).

A análise de variância (ANOVA) e análise de regressão foram, durante muito tempo, o principal esteio da análise e modelagem estatística. Entretanto, estas técnicas têm limitação para lidar com dados desbalanceados e com parentesco entre tratamentos. O método REML permite lidar com essa situação, permitindo maior flexibilidade e eficiência na modelagem. Tal procedimento constitui-se no procedimento padrão para a análise estatística em uma grande gama de aplicações. Em experimentos agronômicos, zootécnicos e florestais, o REML tem substituído com vantagens o método ANOVA criado pelo cientista inglês Ronald Fisher em 1925. Na verdade, o REML é uma generalização da ANOVA para situações mais complexas. Para situações simples, os dois procedimentos são equivalentes, mas para as situações mais complexas encontradas na prática, a ANOVA é um procedimento apenas aproximado. O REML é um método eficiente no estudo das várias fontes de variação associadas à avaliação de experimentos de campo, permitindo desdobrar a variação fenotípica em seus vários componentes genéticos, ambientais e de interação genótipo x ambiente. Estimativas de componentes de variância são essenciais em pelo menos três aplicações: (i) conhecimento do controle genético dos caracteres visando ao delineamento de eficientes estratégias de melhoramento; (ii) predição dos valores genéticos dos candidatos à seleção; (iii) determinação do tamanho de amostra (número de repetições, por exemplo) e forma de amostragem adequada para a estimação precisa de parâmetros e para a maximização da acurácia seletiva.

As principais vantagens práticas do REML/BLUP são: permite comparar indivíduos ou variedades através do tempo (gerações, anos) e espaço (locais, blocos); permite a simultânea correção para os efeitos ambientais, estimação de componentes de variância e predição de valores genéticos; permite lidar com estruturas complexas de dados (medidas repetidas, diferentes anos, locais e delineamentos); pode ser aplicado a dados desbalanceados e a delineamentos não ortogonais; permite utilizar simultaneamente um grande número de informações, provenientes de diferentes gerações, locais e idades, gerando estimativas e predições mais precisas; permite o ajuste de vários modelos alternativos, podendo-se escolher o que se ajusta melhor aos dados e, ao mesmo tempo, é parcimonioso (apresenta menor número de parâmetros).

No caso de dados desbalanceados, a ANOVA conduz a imprecisas estimativas de componentes de variância e conseqüentemente a inacuradas predições de valores genéticos. Um *software* de fácil aplicação prática, destinado ao uso corriqueiro no melhoramento genético é o Selegen-REML/BLUP (Resende, 2002b). As propriedades teóricas desejáveis da metodologia de modelos mistos com vistas à estimação de componentes de variância e à predição de valores genéticos bem como as fórmulas adequadas para as várias situações experimentais são apresentadas com detalhes por Henderson (1984) e Resende (1999; 2000a; 2002 a) e não serão repetidas aqui.

Na análise de modelos mistos com dados desbalanceados, os efeitos do modelo não são testados via testes F tal como se faz no método da análise de variância. Nesse caso, para os efeitos aleatórios, o teste cientificamente recomendado é o teste da razão de verossimilhança (LRT). Para os efeitos fixos, um teste F aproximado pode ser usado. Um quadro similar ao quadro da análise de variância pode ser elaborado. Tal quadro pode ser denominado de Análise de Deviance (ANADEV) e é estabelecido segundo os seguintes passos:

a) Obtenção do ponto de máximo do logaritmo da função de verossimilhança residual (Log L) para modelos com e sem o efeito a ser testado;

a) Obtenção da deviance D = -2 Log L para modelos com e sem o efeito a ser testado;

b) Fazer a diferença entre as deviances para modelos sem e com o efeito a ser testado, obtendo a razão de verossimilhança (LR);

c) Testar, via LRT, a significância dessa diferença usando o teste qui-quadrado com 1 grau de liberdade.

Considere como exemplo o seguinte experimento, conduzido no delineamento de blocos ao acaso com várias plantas por parcela. Tem-se então o seguinte modelo, y = u + g + b + gb + e, em que g refere-se ao efeito aleatório de genótipos, b refere-se ao efeito fixo de blocos, gb refere-se ao efeito aleatório de parcela e e refere-se ao resíduo aleatório dentro de parcela. A seguinte análise de deviance (ANADEV) pode ser realizada.

Efeito	Deviance	LRT(Qui-quadrado[d])	Comp.Var.	Coef. Determ.
Genótipos	647.1794[+]	6.5546**	0.032924*	h2g = 0.0456*
Parcela	654.1289[+]	13.5041**	0.068492**	c2parc = 0.0948**
Resíduo	-	-	0.6206	c2res=0.8595
Modelo Completo	640.6248	-	-	c2total=1.0000
Bloco	-	F = 7.0172**	-	-

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente..
+ Deviance do modelo ajustado sem os referidos efeitos
[d] Distribuição com 1 grau de liberdade.

O *software* Selegen-Reml/Blup (descrito em livro que acompanha esse) fornece as deviances quando se rodam os modelos com ou sem (basta zerar os coeficientes de determinação c2 correspondentes, na tela do Selegen) os efeitos a serem testados. De posse dessas deviances, torna-se fácil construir a tabela da análise de deviance. No presente exemplo, verifica-se que os efeitos de genótipos e de parcelas são significativos. Conseqüentemente, os respectivos componentes de variância são significativamente diferentes de zero, assim como os respectivos coeficientes de determinação (herdabilidade dos efeitos genotípicos – h2g e coeficiente de determinação dos efeitos de parcela -c2parc, conforme obtidos pelos modelos 1 e 2 do Selegen). O fator bloco, considerado de efeito fixo, foi testado via F de Snedecor. A análise de deviância é uma generalização (para os casos balanceado e desbalanceado) da clássica análise de variância.

No Selegen é também apresentado o desvio padrão da herdabilidade individual. De posse da estimativa da herdabilidade e de seu desvio padrão pode-se também inferir sobre a significância dos efeitos genotípicos e, conseqüentemente, sobre a presença de variabilidade genotípica significativa. Essa é outra forma de avaliar a significância dos efeitos genéticos, além daquela já relatada por meio da análise de deviance via o teste da razão de verossimilhança (LRT).

5 MODELO LINEAR MISTO GERAL PARA PREDIÇÃO DE VARIÁVEIS ALEATÓRIAS E ESTIMAÇÃO DE COMPONENTES DE VARIÂNCIA E EFEITOS FIXOS

A seguir é apresentado um modelo geral de estimação REML e predição BLUP com incorporação de um componente espacial.

Um modelo linear misto geral é da forma (Henderson, 1984): $y = Xb + Za + e$ (1), com as seguintes distribuições e estruturas de médias e variâncias:

$a \sim N(0,\ G)$ $\qquad$ $E(y) = Xb$

$e \sim N(0,\ R)$ $\qquad$ $Var(y) = V = ZGZ' + R$

em que:

y: vetor de observações.

b: vetor paramétrico dos efeitos fixos, com matriz de incidência *X*.

a: vetor paramétrico dos efeitos aleatórios, com matriz de incidência Z.

e: vetor de erros aleatórios.

G: matriz de variância – covariância dos efeitos aleatórios.

R: matriz de variância – covariância dos erros aleatórios.

0: vetor nulo.

Assumindo como conhecidos *G* e *R*, a simultânea estimação dos efeitos fixos e predição dos efeitos aleatórios pode ser obtida pelas equações de modelo misto dadas por:

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

A solução deste sistema para $\hat{b}$ e $\hat{a}$ conduz a resultados idênticos aos obtidos por:

$\hat{b} = (X'V^{-1}X)^{-}X'V^{-1}y$: estimador de quadrados mínimos generalizados (GLS), ou melhor estimador linear não viciado (BLUE) de *b*.

$\hat{a} = GZ'V^{-1}(y - X\hat{b}) = C'V^{-1}(y - X\hat{b})$: melhor preditor linear não viciado (BLUP) de *a*; em que *C'* = *GZ'*: matriz de covariância entre a e *y.*

Quando *G* e *R* não são conhecidas, os componentes de variância a elas associados podem ser estimados eficientemente empregando-se o procedimento REML (Patterson e Thompson, 1971). Exceto por uma constante, a função de verossimilhança restrita a ser maximizada é dada por:

$$L = -\frac{1}{2}\ (\log|XV^{-1}X| + \log|V| + v\ \log\ \sigma_e^2 + y'Py/\sigma_e^2)$$

$$= -\frac{1}{2}\ (\log|C^*| + \log|R| + \log|G| + v\ \log\ \sigma_e^2 + y'Py/\sigma_e^2)$$

em que:

$V = R + ZGZ'$; $\quad P = V^{-1} - V^{-1}X\ (X'V^{-1}X)^{-1}X'V^{-1}$;

v = N-r(x): graus de liberdade, em que *N* é o número total de dados e *r(x)* é o posto da matriz *X*;

*C**: matriz dos coeficientes das equações de modelo misto.

A função (*L*) de verossimilhança restrita expressa em termos do logaritmo pode ser maximizada (visando obter as estimativas REML dos componentes de variância) empregando-se diferentes algoritmos tais quais: (i) "E*xpectation – Maximization"* (EM) de Dempster et al. (1977); (ii) "*Derivative Free*" (DF) de Graser et al. (1987); (iii) "*Average Information"* (AI) de Johnson e Thompson (1995) e de Gilmour et al. (1995); (iv) "*Parameter extended EM*" de Foulley e Van Dyk (2000). Estes algoritmos geraram as denominações EMREML, DFREML, AIREML e PX-EM, respectivamente. Detalhes sobre esses métodos numéricos são apresentados no Capítulo 4.

Sendo geral, o modelo (**1**) contempla vários modelos inerentes às diferentes situações, tais quais:

a) Modelo univariado, ajustando apenas o vetor de efeitos aditivos

a: vetor de efeitos genéticos aditivos;

$G = A\sigma_a^2; \quad R = I\sigma_e^2$, em que:

σ_a^2: variância genética aditiva.

A: matriz de correlação genética aditiva entre os indivíduos em avaliação.

σ_e^2: variância residual.

b) Modelo univariado com medidas repetidas, ajustando os efeitos aditivos e de ambiente permanente (p) (Modelo de Repetibilidade)

$$y = Xb + Za + e \qquad Var(a^*) = A\sigma_a^2; \qquad Var(p) = I\sigma_p^2; \qquad R = I\sigma_e^2$$

$$= Xb + Z_1a^* + Z_2p + e, \ em\ que:$$

σ_p^2: variância dos efeitos permanentes.

c) Modelo multivariado, ajustando os efeitos aditivos

No caso bivariado tem-se:

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}; \quad a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix};$$

$$G = A \otimes G_O; \quad R = I \otimes R_O;$$

$$G_O = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{21}} & \sigma^2_{a_2} \end{bmatrix}; \quad R_O = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{21}} & \sigma^2_{e_2} \end{bmatrix} \quad ou \quad R_O = \begin{bmatrix} \sigma^2_{e_1} & 0 \\ 0 & \sigma^2_{e_2} \end{bmatrix}, \ em\ que:$$

$\sigma_{a_{12}}$: covariância genética aditiva entre os caracteres 1 e 2.

$\sigma_{e_{12}}$: covariância residual entre os caracteres 1 e 2.

d) Modelo geoestatístico ou de séries temporais para análise espacial

$R = \Sigma$: matriz não diagonal que considera a correlação entre resíduos, por exemplo, linhas auto-regressivas e colunas auto-regressivas ou estrutura de covariância baseada em semivariâncias ajustadas, para contemplar a autocorrelação espacial entre as observações.

6 BLUP/REML SOB MODELOS INDIVIDUAL, INDIVIDUAL REDUZIDO, GAMÉTICO E DE GENÓTIPOS TOTAIS

6.1 Princípio Básico da Predição de Valores Genéticos

Um modelo estatístico geral para uma observação fenotípica é dado por:

$y = \mu + r + g + e$, em que:

y: valor fenotípico;

μ: média geral;

r: efeitos ambientais identificáveis ou atribuíveis a determinadas causas como repetições, locais, anos, etc ...

g: efeito genético;

e: efeitos ambientais não identificáveis ou completamente aleatórios.

Na predição dos efeitos genéticos (g), o primeiro passo a ser adotado é a correção dos dados observados (y), para os efeitos ambientais identificáveis (r) (van Vleck et al., 1987; Resende, 2002).

Ignorando a média geral, esta correção é dada por: $(y - r) = (g + e)$.

A partir desta correção, a obtenção de g, livre de e, é dada pela multiplicação de (g + e) pela herdabilidade (h^2), ou seja, obtendo-se o preditor $\hat{g} = h^2(y - r) = h^2(g + e) = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}(y - r)$.

A acurácia ($r_{\hat{g}g}$) da predição do valor genético, nesta situação simples, equivale à raiz quadrada da herdabilidade. Assim, quanto maior a herdabilidade, maior é a acurácia seletiva e a precisão na seleção, visto que a variância do erro de predição do valor genético é dada por $PEV = 1 - \frac{r_{\hat{g}g}^2}{\sigma_g^2}$, ou seja, quanto maior a acurácia, menor é a variância do erro de predição. Em situações mais complexas, a acurácia continua sendo função direta da herdabilidade, ou seja, quanto maior a h^2, maior também é $r_{\hat{g}g}$ (van Vleck et al., 1987).

Matricialmente, o modelo geral é dado pelo seguinte modelo linear misto:

$y = Xr + Zg + e$

y: vetor de observações fenotípicas.

r: vetor de efeitos fixos ou efeitos ambientais identificáveis.

g: vetor de efeitos genéticos, aleatório.

e: vetor de efeitos ambientais não identificáveis, aleatório.

X e Z: matrizes de incidência para r e g, respectivamente.

No contexto desse modelo linear misto, a correção dos dados é dada por $y - E(y) = (y - Xr)$. E(y) ou Xr equivalerá à média da repetição (sob um modelo de médias e não de efeitos) quando o efeito da repetição for considerado fixo e equivalerá à média geral quando o efeito de repetição for considerado aleatório.

Entretanto, nas situações práticas, os efeitos ambientais identificáveis não são sempre puramente ambientais, mas, por vezes, incluem também frações de efeitos genéticos, confundidos com os efeitos ambientais. Assim, devem ser classificados como efeitos ambientais identificáveis àqueles predominantemente ambientais, ou seja, aqueles em que as frações de efeitos genéticos confundidos são desprezíveis. E esta classificação nem sempre é trivial.

6.2 Influência da Heterogeneidade Ambiental entre Repetições e Tamanho da Repetição na Eficiência dos Procedimentos Alternativos de Predição

Pela metodologia de modelos mistos, a predição BLUP dos valores genéticos pode ser obtida, alternativamente, considerando o efeito de repetição (r) como fixo ou aleatório. Considerando tal efeito como fixo ou predominantemente ambiental, as equações de modelo misto são dadas por:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}, \text{ em que:} \qquad \textbf{(1)}$$

$$\frac{\sigma_e^2}{\sigma_g^2} = \frac{1-h^2}{h^2}$$

Verifica-se que a predição BLUP, no caso, depende do conhecimento dos componentes de variância σ_g^2 e σ_e^2 ou da herdabilidade $h^2 = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$, que é uma herdabilidade individual dentro da repetição.

Considerando o efeito de repetição como aleatório, as equações de modelo misto são dadas por:

$$\begin{bmatrix} X'X + I\dfrac{\sigma_e^2}{\sigma_r^2} & X'Z \\ Z'X & Z'Z + A^{-1}\dfrac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}, \text{ em que:} \qquad \textbf{(2)}$$

$\dfrac{\sigma_e^2}{\sigma_g^2} = \dfrac{1-h^{2^*}-r^2}{h^{2^*}}$, sendo que h^{2^*} é dada por $h^{2^*} = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 + \sigma_r^2}$, que é uma herdabilidade individual no experimento como um todo e σ_r^2 é a variância entre repetições. Note-se que $h^2 \geq h^{2^*}$, ou seja, h^2 só será igual a h^{2^*} quando $\sigma_r^2 = 0$. O componente r^2 é dado por $r^2 = \dfrac{\sigma_r^2}{\sigma_g^2 + \sigma_e^2 + \sigma_r^2}$.

Verifica-se que a equação **(1)** é um caso particular de **(2)** quando σ_r^2 é muito grande em relação a σ_e^2, ou seja, quando $\sigma_r^2 \to \infty$. Assim, com grande variação ambiental entre repetições, ou seja, com grande heterogeneidade ambiental entre repetições, tem-se:

(i) o modelo **(2)** passa a equivaler ao **(1)** e, portanto, a h^{2*} estimada por REML simultaneamente à predição BLUP é equivalente a h^2;

(ii) os dois modelos conduzem à idênticas predições e acurácias de predição;

(iii) o modelo efetivamente usado é o **(1)**, ou seja, com efeito fixo de repetição.

Quando σ_r^2 é muito pequeno em relação a σ_e^2, ou seja, $\sigma_r^2 \to 0$, não existe heterogeneidade ambiental entre repetições. Neste caso, não existe efeito de repetição e pelo modelo **(2)** pode-se concluir pela não significância de tal efeito, pelo teste da razão de verossimilhança (LRT). Nesta situação, pode-se concluir:

(i) não há necessidade de consideração e/ou ajuste para o efeito de repetição;

(ii) corrigir cada observação pela média do bloco ou pela média geral conduzirá a resultados idênticos;

(iii) $\hat{h}^{2*} = \hat{h}^2$;

(iv) não se justifica usar o modelo **(1),** o qual conduzirá a super-parametrização, ou seja, nesse caso os efeitos de blocos não devem ser tratados como fixos.

Entretanto, na prática, σ_r^2 apresenta magnitude intermediária, de forma que os modelos podem ser efetivamente diferentes e, para comparar os modelos **(1)** e **(2)** é necessário inferir sobre a fração de σ_g^2 que fica confundida no efeito de bloco. Essa fração depende do tamanho de cada repetição ou bloco. Blocos de grande tamanho conduzem a dois fatos: (i) o coeficiente de determinação ambiental da média do bloco tende a 1, ou seja, $r_{ma}^2 = \dfrac{\sigma_r^2}{\sigma_p^2 / f + \sigma_d^2/(nf) + \sigma_r^2}$ tende a 1, em que n é o número de plantas por parcela e f é o número de progênies ou famílias em avaliação; (ii) o coeficiente de determinação genética da média do bloco tende a zero, ou seja,

$r_{mg}^2 = \dfrac{(\sigma_{gd}^2)/(nf)}{\sigma_p^2/f + \sigma_d^2/(nf) + \sigma_r^2}$ tende a zero, em que σ_{gd}^2 refere-se à variação genética dentro de família, σ_p^2 é a variância ambiental entre parcelas e σ_d^2 é a variância total dentro de parcela. Neste caso, não existem informações genéticas a serem recuperadas no efeito de repetições ($r_{mg}^2 = 0$) e os efeitos de repetições podem ser considerados como fixos ($r_{ma}^2 = 1$) sem prejuízo à análise.

Resende e Higa (1994) decompuseram a σ_g^2 total de um teste de famílias de meios irmãos com várias plantas por parcela, em frações retidas nos vários efeitos do modelo estatístico $Y_{ijk} = (Y_{ijk} - \overline{Y}_{ij.}) + (\overline{Y}_{i..} - \overline{Y}_{...}) + (\overline{Y}_{ij.} - \overline{Y}_{i.} - \overline{Y}_{.j.} + \overline{Y}_{...}) + (\overline{Y}_{.j.} - \overline{Y}_{...})$. As várias frações obtidas foram:

(i) Efeito do indivíduo dentro de parcela $(Y_{ijk} - \overline{Y}_{ij.})$:

$((n-1)/n)(1-\rho)\ \sigma_g^2$

(ii) Efeito de progênie $(Y_{i..} - \overline{Y}_{...})$:

$((f-1)/f)\dfrac{1+(nb-1)\ \rho}{nb}\ \sigma_g^2$

(iii) Efeito de parcela $(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})$:

$((f-1)/f)((b-1)/b)\dfrac{1-\rho}{n}\ \sigma_g^2$

(iv) Efeito de bloco $(\overline{Y}_{.j.} - \overline{Y}_{...})$:

$((b-1)/b)\dfrac{1-\rho}{nf}\ \sigma_g^2$

Considerando σ_g^2 a variância genética aditiva, ρ é o coeficiente de correlação genética aditiva entre indivíduos da progênie, ou seja, ρ = 0,25 para meios-irmãos e ρ = 0,5 para irmãos germanos. As quantidades n, f e b referem-se ao número de plantas por parcela, número de famílias

e número de repetições, respectivamente. As notações Y_{ijk} , $\overline{Y}_{ij.}$, $\overline{Y}_{i.}$, $\overline{Y}_{.j.}$ e $\overline{Y}_{...}$ referem-se aos valores individuais, média de parcela, média de família, média de bloco e média geral, respectivamente.

Verifica-se que a fração da variação genética retida no efeito de repetição depende inversamente do tamanho da repetição (N = nf). Com f = 100 famílias e n = 5 plantas por parcela, as frações da variação genética retida serão 0,75/500 = 0,0015 para progênies de meios irmãos e 0,50/500 = 0,0010 para progênies de irmãos germanos. Estas frações são desprezíveis e, portanto, neste caso, não existem efeitos genéticos a serem recuperados nos efeitos de repetições e os efeitos de repetições podem ser considerados como fixos desde que a variação entre repetições (σ_r^2) não tenda a zero.

Com uma planta por parcela, o efeito de repetições retém $\frac{1-\rho}{f}\ \sigma_g^2$. Neste caso, com as mesmas 100 progênies as frações de variância genética retida serão 0,75/100 = 0,0075 para progênies de meios irmãos e 0,50/100 = 0,0050 para progênies de irmãos-germanos. Estas frações representam menos de 1 % da variação genética total e são igualmente desprezíveis. Com teste clonal ou teste de híbridos simples ou teste de linhagens em vez de teste de progênies, tem-se ρ = 1 e nenhuma variação genética será retida entre repetições.

Mesmo que a sobrevivência não seja 100 %, estas inferências são válidas, visto que as repetições são altamente conectadas, com conexão próxima a 100 %, conforme conceito de conectabilidade de Weeks e Williams (1964). Com sobrevivência muito baixa e especialmente nos delineamentos com uma planta por parcela, as repetições tenderão a ser incompletas e caracterizarão blocos incompletos. Nesta situação, os blocos incompletos são melhor modelados como efeitos aleatórios (visando à recuperação da informação genética entre famílias inter-blocos), sendo a mesma afirmativa válida para os delineamentos em látice (Resende e Fernandes, 2000).

Resende e Fernandes (1999) demonstraram a equivalência entre índices multi-efeitos ótimos derivados para o caso balanceado por Resende e Higa (1994) e suas correspondentes predições BLUP via metodologia de modelos mistos. Assim, índices multi-efeitos incluindo todos os efeitos aleatórios do modelo estatístico podem ser utilizados para derivar os seus correspondentes BLUP,

para vários modelos alternativos. Para o modelo **(1)**, porém para o caso de delineamentos com várias plantas por parcela (fato que requer a expansão do modelo **(1)** para incluir o efeito de parcela), o correspondente índice multi-efeitos é dado por (Resende, 1991; Resende & Higa, 1994):

$$\hat{g}_1 = b_1\ (Y_{ijk} - \overline{Y}_{ij.}) + b_2\ (\overline{Y}_{i..} - \overline{Y}_{...}) + b_3\ (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})$$

Para o modelo **(2)** expandido, o correspondente índice multi-efeitos é dado por:

$$\hat{g}_2 = b_1\ (Y_{ijk} - \overline{Y}_{ij.}) + b_2\ (\overline{Y}_{i..} - \overline{Y}_{...}) + b_3\ (\overline{Y}_{ij.} - \overline{Y}_{i.} - \overline{Y}_{.j.} + \overline{Y}_{...}) + b_4\ (\overline{Y}_{.j.} - \overline{Y}_{...})$$

Os ponderadores dos vários efeitos aleatórios dos índices multi-efeitos são:

$b_1 = h_d^2 = \dfrac{(1-\rho)\ \sigma_g^2}{\sigma_d^2}$: herdabilidade do efeito de indivíduo dentro de parcela;

$b_2 = h_f^2 = \dfrac{\dfrac{1+(nb-1)\ \rho}{nb}\ \sigma_g^2}{\sigma_f^2 + \sigma_p^2/b + \sigma_d^2/nb}$: herdabilidade do efeito de progênie;

$b_3 = h_p^2 = \dfrac{\dfrac{1-\rho}{n}\ \sigma_g^2}{\sigma_d^2/n + \sigma_p^2}$: herdabilidade do efeito de parcela;

$b_4 = h_r^2 = \dfrac{\dfrac{1-\rho}{nf}\ \sigma_g^2}{\sigma_r^2 + \sigma_p^2/f + \sigma_d^2/nf}$: herdabilidade do efeito de bloco, em que:

σ_f^2, σ_p^2, σ_d^2 *e* σ_r^2: variância entre famílias, entre parcelas, dentro de parcelas e entre repetições, respectivamente.

No caso de uma planta por parcela, os índices multi-efeitos são:

$\hat{g}_1 = b_5\ (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}) + b_6\ (\overline{Y}_{i.} - \overline{Y}_{..})$, que equivale exatamente ao BLUP pelo modelo **(1)** para o caso balanceado.

$\hat{g}_2 = b_5\ (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}) + b_6\ (\overline{Y}_{i.} - \overline{Y}_{..}) + b_7\ (\overline{Y}_{.j} - \overline{Y}_{..})$, que equivale exatamente

ao BLUP pelo modelo **(2)** para o caso balanceado.

Os ponderadores dos vários efeitos são:

$b_5 = \dfrac{(1-\rho)\ \sigma_g^2}{\sigma^2}$: herdabilidade dentro de progênie;

$b_6 = \dfrac{\dfrac{1+(b-1)\ \rho}{b}\ \sigma_g^2}{\sigma_f^2 + \sigma^2/b}$: herdabilidade entre progênies;

$b_7 = \dfrac{\dfrac{1-\rho}{p}\ \sigma_g^2}{\sigma_r^2 + \sigma^2/f}$: herdabilidade do efeito de bloco, em que:

σ^2 : variância residual no delineamento com uma planta por parcela.

Conforme esperado, verifica-se que quando σ_r^2 é muito grande ($\sigma_r^2 \to \infty$), $\hat{g}_2$ converge para $\hat{g}_1$ (pois $b_4 = 0$ e $b_7 = 0$), assim como o modelo **(2)** converge para o modelo **(1)** nas equações de modelo misto para a predição BLUP. Adicionalmente, esta convergência também ocorre quando o tamanho da repetição é grande (nf alto), independentemente da magnitude de σ_r^2, desde que o mesmo não seja zero. Isto porque o numerador de b_4 e b_7 tenderia a zero (e não haveria informação genética a ser recuperada no efeito de repetição). Este resultado foi confirmado por um estudo de simulação realizado por Ugarte et al. (1992), que concluíram que um modelo do tipo BLUP(2) é mais vantajoso que um modelo do tipo BLUP(1) somente quando cada nível do efeito em consideração (repetição) for pequeno e, simultaneamente, σ_e^2/σ_r^2 for de alta magnitude, ou seja, quando existe pequena heterogeneidade ambiental entre repetições.

Em uma situação real de campo, na experimentação com espécies perenes, os fatores repetição grande (nf alto) e grande heterogeneidade ambiental entre repetições (σ_r^2 de grande magnitude) contribuem simultaneamente para que b_4 e b_7 tendam a zero e, portanto, para que a predição pelo modelo **(1)** (ou pelo índice $\hat{g}_1$) seja adequada.

Do exposto até aqui, conclui-se que considerar os efeitos de repetições como aleatórios é vantajoso quando o tamanho da repetição for pequeno e, simultaneamente, σ_e^2 / σ_r^2 for de alta magnitude, ou seja, quando existe pequena heterogeneidade ambiental entre repetições. Nas demais situações, considerar como fixos ou aleatórios conduz à mesma eficiência. Neste último caso, as eficiências dos dois procedimentos alternativos não podem ser julgadas apenas pelos denominadores de h^{2*} e h^2. Deve ser considerado o conceito de herdabilidade ajustada (Resende e Thompson, 2003), a qual é dada por $h_{aj}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. Verifica-se que nesse caso, as herdabilidades ajustadas obtidas pelos dois modelos tenderão a ser equivalentes. A herdabilidade ajustada refere-se a uma herdabilidade livre de todos os efeitos ambientais aleatórios ajustados no modelo e, portanto, contempla em seu denominador apenas os componentes de variância genotípica e residual. Devido a essa composição, a herdabilidade ajustada permite comparar modelos alternativos de análise pois é função da variância residual peculiar ao ajuste de cada modelo (quanto menor a variação residual melhor o modelo) e também inclui a quantidade de variação genética recuperada pelo modelo de análise. Tal herdabilidade é livre da flutuação nos demais componentes de variância relativos aos efeitos ambientais, pois é inversamente proporcional apenas ao erro ou variação aleatória residual não ajustada no modelo. A herdabilidade ajustada está associada ao fator de *shrinkage* para os efeitos genotípicos nas equações de modelo misto, pois $\lambda_1 = \frac{1 - h_{aj}^2}{h_{aj}^2}$, mesmo para modelos com vários efeitos aleatórios, além do erro. Assim, informa sobre a confiabilidade dos valores fenotípicos ajustados para todos os efeitos fixos e demais efeitos aleatórios do modelo como indicadores dos efeitos genotípicos verdadeiros. O modelo que propicia maior confiabilidade de seus valores fenotípicos ajustados para todos os demais efeitos do modelo é o melhor modelo.

Um outro fator a considerar na escolha entre efeitos fixos ou aleatórios de repetições refere-se ao vício na comparação genética de indivíduos através das repetições. A consideração dos efeitos de repetição como fixos propicia predições invariantes aos efeitos fixos, removendo assim vícios nas comparações genéticas através das repetições. Tal consideração é, sobretudo, importante quando existe uma alocação não aleatória dos materiais genéticos nas repetições e, neste caso, o BLUP minimiza o vício causado por associações entre efeitos de repetições e o nível

genético dos materiais que nelas se desenvolvem. Por outro lado, quando não existem associações entre efeitos de repetições e o nível genético dos materiais, não existem vícios (PEV aproximadamente igual ao erro quadrático médio) mesmo quando se consideram como aleatórios os efeitos de repetições.

Considerando que o delineamento em blocos completos balanceados propicia uma completa conectabilidade (capacidade para estimar diferenças ambientais entre repetições, conforme Weeks e Williams, 1964), a adoção do modelo com efeito fixo de repetição não é necessário quando o objetivo é a seleção de genótipos (tratamentos) com base em seus comportamentos através das repetições. Por outro lado, quando o objetivo é a seleção dos indivíduos componentes das progênies, os quais são únicos e não repetidos, o tipo de alocação dos materiais genéticos nas repetições, torna-se então relevante. Na prática do melhoramento florestal, em geral, as repetições são estabelecidas com mudas de padrão similar de crescimento, ou seja, as melhores mudas são estabelecidas em uma mesma repetição, o mesmo acontecendo com as piores. Isto caracteriza uma alocação não aleatória de materiais genéticos nas repetições. Em situações de alocação não aleatória dos indivíduos nas repetições, é melhor tratar os efeitos de repetição como fixos para propósitos práticos de avaliação genética.

A grande questão envolvida na comparação não viciada de indivíduos através das repetições é o confundimento entre efeitos genéticos e ambientais na média da repetição. Com alocação não aleatória, torna-se mais difícil o ajuste para os efeitos ambientais e a separação entre os efeitos genéticos e ambientais confundidos. Nesta situação, os efeitos de repetição devem ser tratados como fixos para que se obtenha o BLUP. Segundo Foulley et al. (1990), uma simples razão para explicar isto, refere-se ao fato do BLUP ser invariante à translação nos efeitos considerados fixos. Isto pode ser explicado também invocando um argumento Bayesiano, de forma que o BLUP com r fixo equivale a um estimador Bayesiano usando uma distribuição *a priori* não informativa para r.

Em resumo, a comparação entre os modelos BLUP(1) e BLUP(2) depende de três fatores: (*i*) do tamanho da repetição; (*ii*) do tipo de alocação dos materiais genéticos nas repetições (aleatória ou não); (*iii*) da relação σ_e^2 / σ_r^2.

Considerar efeitos de repetições completas (balanceadas ou não em termos do número de plantas por progênie) como fixos ou aleatórios é um compromisso entre garantir não vício (o U do BLUP) ou recuperar informação genética retida no referido efeito, respectivamente. Considerar como fixo garante não vício, mas, pode perder informação genética quando o tamanho da repetição é muito pequeno. Considerar aleatório garante o uso de toda informação genética, mas não garante efetivamente o não vício.

Neste sentido, o tamanho da repetição é de grande importância na decisão. Na situação de pequeno tamanho de repetição, a diferença nos valores de acurácia associados ao BLUP(1) e BLUP(2) é devido a dois fatores: (*i*) aumento na acurácia do BLUP(2) por causa do uso adicional do efeito de repetição e (*ii*) redução na acurácia do BLUP(1) em função da estimação do efeito de repetição, a partir de um menor número de observações do que aquele usado na estimação do efeito da média geral pelo BLUP(2).

A estimação dos efeitos fixos simultaneamente à predição dos valores genéticos conduz a um acréscimo (em relação ao BLP que supõe os efeitos fixos como conhecidos) na PEV, porque um erro adicional é introduzido em função do uso de uma estimativa destes efeitos e não de seu valor verdadeiro. A PEV pode ser expressa por $PEV = G - C'V^{-1}C + C'V^{-1}X\,(X'V^{-1}X)^{-}\,X'V^{-1}C$ (em que $G = Var(a)$ e $C = COV\,(a, y)$) (Henderson, 1984) em que a terceira parte desta expressão refere-se ao acréscimo devido à estimação dos efeitos fixos. Verifica-se, assim, que a contribuição para a PEV, associada à estimação dos efeitos fixos, é maior para o BLUP(1) do que para o BLUP(2). Este resultado é devido ao menor número de observações usados na estimação das médias de repetição pelo BLUP(1). De maneira geral, na medida em que o número de observações em cada nível do efeito fixo aumenta, diminui o acréscimo na PEV (Kennedy e Trus, 1993). Com o aumento do número de tratamentos, tanto a unidade repetição quanto experimento aumentam e, conseqüentemente, isto conduz a uma redução na contribuição da estimação dos efeitos fixos para a PEV, tanto para o BLUP(1) quanto para o BLUP(2).

Considerando o efeito de repetição como aleatório, o modelo não será mais misto e sim aleatório, segundo a classificação tradicional dos modelos em fixos, aleatórios e mistos (Henderson, 1953; Searle, 1971). sob um modelo aleatório (tendo somente a média geral μ como efeito fixo), o fato de considerar μ como conhecido ou estimá-lo a partir dos dados, conduz aos mesmos

resultados práticos (Visscher e Goddard, 1993). Neste caso, o BLUP e o BLP (melhor preditor linear) são equivalentes para propósitos de seleção. Também neste caso, em termos de estimação de componentes de variância, os métodos de máxima verossimilhança (ML) e REML se equivalem, pois a divisão por η ou por (η - 1) nos estimadores da variância residual por ML e REML, respectivamente, não conduzem a diferenças práticas nas estimativas (Blasco, 2001). Em resumo, no modelo aleatório, o procedimento ML/BLP é adequado para a estimação/predição.

Uma outra questão a considerar na classificação entre efeito fixo e aleatório de repetições refere-se a questão do estimador de mínimo erro quadrático médio, conforme discutido no tópico 2. Nesse aspecto, concluiu-se que quando o número de repetições é maior que dez, é melhor tratar esses efeitos como aleatórios. Então, conclui-se, finalmente:

(i) Pode-se tratar os efeitos de repetições (blocos) como aleatórios quando o numero de repetições é maior que dez e, simultaneamente, a alocação dos indivíduos nas repetições é aleatória. Isto vale para blocos completos ou incompletos e independe do tamanho da repetição;

(ii) É melhor tratar os efeitos de repetições (blocos) como fixos quando o número de repetições é menor que dez, desde que o tamanho da repetição seja grande e a variação entre repetições não tenda a zero;

(iii) Com número de repetições menor que dez e variação entre repetições tendendo a zero, deve-se ignorar os efeitos de blocos;

(iv) Com número de repetições menor que dez, variação entre repetições não tendendo a zero e tamanho da repetição pequeno (pequeno número de tratamentos), deve-se usar os estimadores de James-Stein para o referido efeito. Nessa situação não haverá recuperação de informação genética no efeito de repetições. Isto não é relevante quando o interesse é a seleção de tratamentos avaliados em blocos completos. Nota-se também que poderá haver insuficientes graus de liberdade para o resíduo, se tanto o número de tratamentos quanto de repetições forem muito baixos, sendo que é melhor evitar essa situação (graus de liberdade do resíduo menor que oito) por ocasião do planejamento da experimentação.

Para verificar a magnitude da variação entre repetições, pode-se, inicialmente, adotar o modelo aleatório e, então, verificar se o modelo com efeitos fixos de repetições é mais adequado ou não. Tratar os efeitos de repetição como fixos não limita as inferências dos resultados para o conjunto de repetições analisado. O importante é que os efeitos da interação efeitos genéticos x efeitos ambientais sejam de natureza aleatória e isto ocorrerá sempre que os efeitos genéticos forem tratados como aleatórios. Nesse caso, os efeitos genéticos (g) preditos podem ser extrapolados para quaisquer ambientes, bastando realizar a soma $\hat{g} + E(y)$, que fornece o valor genético predito para um ambiente com média E(y).

6.3. Modelo Individual

Para o caso de resíduos não correlacionados, são apresentados a seguir exemplos de avaliação genética pelo procedimento BLUP Individual.

Considere o seguinte conjunto de dados e genealogia, associados à avaliação de uma espécie florestal, através de famílias de meios-irmãos, em que cada duas árvores constituem uma parcela.

Indivíduo	Família (Mãe)	*Bloco*	Parcela	Árvore	Diâmetro (cm)
4	1	1	1	1	9,87
5	1	1	1	2	14,48
6	2	1	2	1	8,91
7	2	1	2	2	14,64
8	3	1	3	1	9,55
9	3	1	3	2	7,96
10	1	2	4	1	16,07
11	1	2	4	2	14,01
12	2	2	5	1	7,96
13	2	2	5	2	21,17
14	3	2	6	1	10,19
15	3	2	6	2	9,23

Nesta situação, o modelo linear misto (considerando os efeitos de blocos como fixos) adequado para a descrição dos dados equivale a:

y = Xb + Za + Wc + e, em que:

y , b, a, c e e: vetores de dados, de efeitos fixos (médias de blocos), de efeitos aditivos (aleatório), de efeitos de parcelas (efeitos aleatórios de ambiente comum das parcelas) e de erros aleatórios, respectivamente.

X, Z e W: são matrizes de incidência conhecidas, formadas por valores 0 e 1, às quais associam as incógnitas b, a e c ao vetor de dados y, respectivamente.

A metodologia de modelos mistos permite estimar b pelo procedimento de quadrados mínimos generalizados (GLS) e predizer a e c pelo procedimento BLUP. Para a obtenção destas soluções, basta resolver o seguinte sistema de equações lineares, o qual é denominado equações de modelo misto (EMM):

$$\begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$, em que:

$$\lambda_1 = \frac{1 - h^2 - c^2}{h^2}; \qquad \lambda_2 = \frac{1 - h^2 - c^2}{c^2}$$

A e I = matrizes de parentesco genético aditivo e matriz identidade de ordem apropriada aos dados, respectivamente.

Empregando-se os dados apresentados anteriormente, têm-se as seguintes matrizes de incidência:

$$X_{(12x2)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad Z_{(12x15)} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad W_{(12x6)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

A matriz I no caso é de ordem 6 e a matriz A com a inclusão das 3 matrizes equivale a:

$$A = \begin{bmatrix} 1.0000 & 0 & 0 & 0.5000 & 0.5000 & 0 & 0 & 0 & 0 & 0.5000 & 0.5000 & 0 & 0 & 0 & 0 \\ 0 & 1.0000 & 0 & 0 & 0 & 0.5000 & 0.5000 & 0 & 0 & 0 & 0 & 0.5000 & 0.5000 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 & 0 & 0 & 0 & 0.5000 & 0.5000 & 0 & 0 & 0 & 0 & 5.000 & 5.000 \\ 0.5000 & 0 & 0 & 1.0000 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 \\ 0.5000 & 0 & 0 & 0.2500 & 1.0000 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 \\ 0 & 0.5000 & 0 & 0 & 0 & 1.0000 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 \\ 0 & 0.5000 & 0 & 0 & 0 & 0.2500 & 1.0000 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 \\ 0 & 0 & 0.5000 & 0 & 0 & 0 & 0 & 1.0000 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 \\ 0 & 0 & 0.5000 & 0 & 0 & 0 & 0 & 0.2500 & 1.0000 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 \\ 0.5000 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 1.0000 & 0.2500 & 0 & 0 & 0 & 0 \\ 0.5000 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 1.0000 & 0 & 0 & 0 & 0 \\ 0 & 0.5000 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 1.0000 & 0.2500 & 0 & 0 \\ 0 & 0.5000 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 1.0000 & 0 & 0 \\ 0 & 0 & 0.5000 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 1.0000 & 0.2500 \\ 0 & 0 & 0.5000 & 0 & 0 & 0 & 0 & 0.2500 & 0.2500 & 0 & 0 & 0 & 0 & 0.2500 & 1.0000 \end{bmatrix}$$

Considerando os parâmetros herdabilidade individual no sentido restrito e coeficiente de determinação dos efeitos de parcela, respectivamente, iguais a $h^2 = 0{,}1635 \;\; e \;\; c^2 = 0{,}0779$, obtém-se λ_1 = 4,639 e λ_2 = 9,738. Resolvendo as MME, obtém-se o vetor de soluções:

$$\begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} 10{,}9017 \\ 13{,}1050 \\ 0{,}4370 \\ 0{,}3178 \\ -0{,}7548 \\ 0{,}0224 \\ 0{,}6641 \\ -0{,}1554 \\ 0{,}6422 \\ -0{,}4760 \\ -0{,}6973 \\ 0{,}5649 \\ 0{,}2782 \\ -0{,}6065 \\ 1{,}2321 \\ -0{,}6675 \\ -0{,}8011 \\ 0{,}1585 \\ 0{,}1073 \\ -0{,}2658 \\ 0{,}2579 \\ 0{,}1955 \\ -0{,}4534 \end{bmatrix}$$

No vetor de soluções, os dois primeiros valores referem-se às estimativas BLUE (melhor estimativa linear não viciada) das médias dos blocos 1 e 2, respectivamente; os três valores seguintes são relativos aos efeitos genéticos aditivos preditos para as 3 matrizes; os 12 valores seguintes dizem respeito aos efeitos genéticos aditivos preditos das progênies (descendentes) e os 6 últimos valores, aos efeitos ambientais preditos para as parcelas.

6.4 Modelo Individual Reduzido – MIR

No modelo reduzido, o mesmo sistema de equações se aplica, porém, o vetor de efeitos genéticos aditivos contempla apenas os genitores. No caso em que não se têm observações (dados) referentes a genitoras, as EMM se reduzem a:

$$\begin{bmatrix} \hat{b} \\ \hat{a}_g \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'X & X'Z_1 & X'W \\ Z_1'X & Z_1'Z_1 + A_g^{-1}\lambda_1 & Z_1'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z_1'y \\ W'y \end{bmatrix}$$, em que:

$\lambda_1 = \dfrac{\sigma_e^2 + (3/4)\sigma_a^2}{\sigma_a^2} = \dfrac{4 - h^2 - 4c^2}{4h^2}$ = 5,389755 no presente exemplo;

$\lambda_2 = \dfrac{\sigma_e^2 + (3/4)\sigma_a^2}{\sigma_c^2} = \dfrac{4 - h^2 - 4c^2}{4c^2}$ = 11,31226 no presente exemplo.

As matrizes X, W e y são exatamente como especificado no tópico anterior. A matriz A, no caso de testes de progênie de genitoras não aparentadas, é uma matriz identidade (neste exemplo, de ordem 3).

A matriz Z_1 é composta por valores 0 e 0,5, equivalendo a:

$$Z_{1(12x3)} = \begin{bmatrix} 0{,}5 & 0 & 0 \\ 0{,}5 & 0 & 0 \\ 0 & 0{,}5 & 0 \\ 0 & 0{,}5 & 0 \\ 0 & 0 & 0{,}5 \\ 0 & 0 & 0{,}5 \\ 0{,}5 & 0 & 0 \\ 0{,}5 & 0 & 0 \\ 0 & 0{,}5 & 0 \\ 0 & 0{,}5 & 0 \\ 0 & 0 & 0{,}5 \\ 0 & 0 & 0{,}5 \end{bmatrix},$$

O vetor de soluções deste exemplo é equivalente a:

$$\begin{bmatrix} \hat{b} \\ \hat{a}_g \\ \hat{c} \end{bmatrix} = \begin{bmatrix} 10{,}9017 \\ 13{,}1050 \\ 0{,}4370 \\ 0{,}3178 \\ -0{,}7548 \\ 0{,}1585 \\ 0{,}1073 \\ -0{,}2658 \\ 0{,}2579 \\ 0{,}1955 \\ -0{,}4534 \end{bmatrix}$$

Verifica-se que esse modelo fornece como resultados os valores genéticos integrais das genitoras bem como os efeitos ambientais de bloco e parcelas, exatamente como no modelo MI.

Para cômputo dos valores genéticos dos indivíduos, deve-se empregar a expressão $\hat{a} = Z_1\hat{a}_g + h_d^2(y - X\hat{b} - Z_1\hat{a}_g - W\hat{c})$, a qual fornece os seguintes resultados, que são idênticos aos obtidos pelo MI:

$$\hat{a} = \begin{bmatrix} 0{,}0224 \\ 0{,}6641 \\ -0{,}1554 \\ 0{,}6422 \\ -0{,}4760 \\ -0{,}6973 \\ 0{,}5649 \\ 0{,}2782 \\ -0{,}6065 \\ 1{,}2321 \\ -0{,}6675 \\ -0{,}8011 \end{bmatrix}$$

A herdabilidade dentro de progênies de meios-irmãos para um modelo com parcelas de várias plantas é dada por h_d^2 = (3/4 h^2)/[(3/4) h^2 + (1 – h^2 – c^2)] e equivale a 0,13915 no presente caso. O valor individual corrigido para os efeitos de bloco, parcela e progênie é ponderado pela própria h_d^2 .

As soluções propiciadas pela expressão $\hat{a} = Z_1\hat{a}_g + h_d^2(y - X\hat{b} - Z_1\hat{a}_g - W\hat{c})$ são equivalentes às obtidas pelo método do índice multiefeitos (Resende e Higa, 1994), o qual foi implementado no *software* selegen (Resende et al., 1994). De fato, quando os dados são balanceados, os métodos BLUP e índice multiefeitos são idênticos, ou seja, nesta situação, o método do índice multiefeitos é BLUP.

Como exemplo, o resultado obtido para o primeiro indivíduo é dado por $\hat{a} = 0{,}5 * 0{,}4370 + 0{,}13915\,(9{,}87 - 10{,}9017 - 0{,}5 * 0{,}4370 - 0{,}1585) = 0{,}0224$.

Esta formulação é também benéfica à estimação REML de componentes de variância. Neste caso, os seguintes estimadores REML podem ser usados:

$$\hat{\sigma}^2 = [y'y - \hat{b}'X'y - \hat{a}'_g Z_1'y - \hat{c}'W'y]/[N - r(x)] = \hat{\sigma}_e^2 + (3/4)\hat{\sigma}_a^2$$

$$\hat{\sigma}_a^2 = [\hat{a}'_g A^{-1}\hat{a}_g + \hat{\sigma}^2\ tr\ C^{22}]/q$$

$\hat{\sigma}_c^2 = [\hat{c}'c + \hat{\sigma}^2\ tr\ C^{33}]/s$, em que:

tr: operador traço matricial; r(x): posto da matriz X; q : número de genitoras;

s : número de parcelas; N : número total de dados; C^{22} e C^{33} provêm de:

$$C = \begin{bmatrix} C^{11} & C^{12} & C^{13} \\ C^{21} & C^{22} & C^{23} \\ C^{31} & C^{32} & C^{33} \\ \end{bmatrix}$$ = inversa generalizada da matriz dos coeficientes das EMM.

Para o caso de testes de progênies de irmãos germanos com várias plantas por parcela, a mesma formulação do MIR pode ser aplicada, bastando considerar que:

$$\lambda_1 = \frac{\sigma_e^2 + (1/2)\sigma_a^2}{\sigma_a^2} = \frac{2 - h^2 - 2c^2}{2h^2};$$

$$\lambda_2 = \frac{\sigma_e^2 + (1/2)\sigma_a^2}{\sigma_c^2} = \frac{2 - h^2 - 2c^2}{2c^2};$$

$$h_d^2 = (1/2\ h^2)/[(1/2)\ h^2 + (1 - h^2 - c^2)]\,.$$

6.5 Modelo Gamético – MG

No modelo gamético, o mesmo sistema de equações se aplica, porém, o vetor de efeitos genéticos aditivos contempla apenas metade do valor genético dos genitores. As EMM são:

$$\begin{bmatrix} \hat{b} \\ (\hat{a}/2) \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + A_g^{-1}\lambda_1 & Z'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$, em que:

$\lambda_1 = \dfrac{\sigma_e^2 + (3/4)\sigma_a^2}{(1/4)\sigma_a^2} = \dfrac{4 - h^2 - 4c^2}{h^2}$ = 21,5590 no presente exemplo;

$\lambda_2 = \dfrac{\sigma_e^2 + (3/4)\sigma_a^2}{\sigma_c^2} = \dfrac{4 - h^2 - 4c^2}{4c^2}$ = 11,31226 no presente exemplo.

As matrizes X, W e y são exatamente como especificado no tópico anterior. A matriz A, no caso de testes de progênie de genitoras não aparentadas, é uma matriz identidade (neste exemplo, de ordem 3).

A matriz Z_1 é composta por valores 0 e 1, equivalendo a:

$$Z_{1(12x3)} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

O vetor de soluções deste exemplo é equivalente a:

$$\begin{bmatrix} \hat{b} \\ (\hat{a}/2) \\ \hat{c} \end{bmatrix} = \begin{bmatrix} 10{,}9017 \\ 13{,}1050 \\ 0{,}2185 \\ 0{,}1589 \\ -0{,}3774 \\ 0{,}1585 \\ 0{,}1073 \\ -0{,}2658 \\ 0{,}2579 \\ 0{,}1955 \\ -0{,}4534 \end{bmatrix}$$

Verifica-se que este modelo fornece como resultados a metade dos valores genéticos das genitoras bem como os efeitos ambientais de bloco e parcelas. A variância de (a/2) contempla (1/4) da variância genética aditiva, que equivale à variância genética entre progênies de meios-irmãos.

Para cômputo dos valores genéticos dos indivíduos, deve-se empregar a expressão $\hat{a} = Z(\hat{a}/2) + h_a^2(y - X\hat{b} - Z(\hat{a}/2) - W\hat{c})$, a qual fornece resultados idênticos aos obtidos pelo MI e MIR.

6.6 Modelo de Genótipos Totais – MGT

Considere a avaliação de genótipos (clones, linhagens, híbridos, populações, acessos) não aparentados no delineamento de blocos ao acaso com várias plantas por parcela e uma medição por indivíduo.

Neste caso, o modelo linear misto adequado à descrição dos dados equivale a:

y = Xb + Zg + Wc + e, em que:

y , b, g, c e e: vetores de dados, de efeitos fixos (médias de blocos), de efeitos genotípicos (aleatórios), de efeitos de parcelas (aleatórios) e de erros aleatórios, respectivamente.

X, Z e W : matrizes de incidência para b, g e c, respectivamente.

As equações de modelo misto para a estimação dos efeitos fixos e predição dos efeitos aleatórios equivalem a:

$$\begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + I\lambda_1 & Z'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$, em que:

$$\lambda_1 = \frac{1 - h_g^2 - c^2}{h_g^2}; \qquad \lambda_2 = \frac{1 - h_g^2 - c^2}{c^2}$$

Usando o mesmo conjunto de dados apresentados no tópico 6.3 e os parâmetros $h_g^2 = 0{,}2453$ e $c^2 = 0{,}0779$, tem-se $\lambda_1 = 2{,}75907$; $\lambda_2 = 8{,}68806$ e obtêm-se os efeitos genotípicos preditos equivalentes a 0,8678; 0,6311 e –1,4989, para os clones 1, 2 e 3, respectivamente. No caso, as matrizes X, Z e W são exatamente iguais às do modelo gamético, e Var (g) contempla a variância genotípica total. O parâmetro h_g^2 refere-se à herdabilidade individual no sentido amplo.

7 USO DO REML/BLUP NA AVALIAÇÃO DE TRATAMENTOS GENÉTICOS SOB DIFERENTES DELINEAMENTOS EXPERIMENTAIS E DE CRUZAMENTO

Conforme relatado no tópico anterior, os efeitos de tratamentos genéticos devem ser considerados preferencialmente como aleatórios. Assim, os preditores BLUP e estimadores REML são usados nas várias situações associadas aos diferentes delineamentos experimentais tais como blocos ao acaso, látice, linha e coluna, blocos aumentados. São usados também associados aos vários delineamentos de cruzamentos tais como progênies de polinização aberta, cruzamentos dialélicos, fatoriais, hierárquicos, testes clonais. O BLUP também tem sido importante na predição de híbridos simples de milho e de outras espécies. Os delineamentos mais complexos como os dialélicos e fatoriais permitem uma grande gama de inferências tais quais aquelas sobre os efeitos aditivos (capacidade geral de combinação), efeitos de dominância e da capacidade específica de combinação bem como suas variâncias, tanto em nível individual quanto de famílias. Todas essas situações foram abordadas em detalhes por Resende (1999; 2000a e 2002a) e não serão repetidas aqui. Serão apresentados apenas quatro tópicos complementares.

7.1 BLUP na Seleção em Plantas Autógamas

Os delineamentos de cruzamento associados a autofecundações podem também ser analisados via REML/BLUP. Em espécies autógamas é comum a avaliação, em várias gerações de autofecundação, de linhagens obtidas a partir de cruzamentos entre dois genitores divergentes. Em

vários programas de melhoramento nessas espécies, adota-se alguma forma de seleção precoce na geração F_3, explorando-se a grande variabilidade genética entre e dentro de linhagens F_3. Tal variabilidade contempla 1,5 vezes a variância genética aditiva (σ_a^2), sendo que $0.5\sigma_a^2$ encontra-se dentro de linhagem e 1,0 σ_a^2 encontra-se entre linhagens. Assim, tal geração é adequada para seleção, pois 75 % (1,5) da variação aditiva total (2 σ_a^2) que estará disponível em F_∞ já se encontra disponível em F_3. Dessa forma, a seleção em F_3 por meio de um método preciso como o BLUP é relevante.

Em algumas espécies autógamas como o café arábica e a aveia tem sido realizadas avaliações de plantas individuais em linhagens F_3 (Federizzi et al., 1999). O BLUP para seleção neste caso pode ser derivado, considerando o delineamento de blocos ao acaso com várias plantas por parcela. Usando o índice multi-efeitos derivado por Resende e Higa (1994), tem-se que o índice ótimo ou BLUP para o caso balanceado e, considerando blocos como efeitos fixos, nesse caso, é dado por:

$$I = b_1\delta_{ijk} + b_2 g_i + b_3\ c_{ij}$$

$$= b_1\left(Y_{ijk} - \overline{Y}_{ij}.\right) + b_2\left(\overline{Y}_{i..} - \overline{Y}\ldots\right) + b_3\left(\overline{Y}_{ij}. - \overline{Y}_{i}.. - \overline{Y}_{.j}. + \overline{Y}\ldots\right)$$

A versão matricial deste índice, usando equações de modelo misto, fornece o BLUP generalizado para os casos balanceado e desbalanceado. Os coeficientes do índice, no caso de linhagens F_3 advindas do cruzamento entre dois genitores para gerar a F_1, são dados por:

$b_1 = \dfrac{(1/2)\ \sigma_a^2}{\sigma_\delta^2}$: herdabilidade do efeito de indivíduo dentro de parcela.

$b_2 = \dfrac{(2nb+1)}{2nb}\sigma_a^2$: herdabilidade do efeito de família.

$\sigma_f^2 + \sigma_p^2/b + \sigma_\delta^2/nb$

$b_3 = \dfrac{[(1/2)/n]\ \sigma_a^2}{\sigma_c^2 + \sigma_\delta^2/n}$: herdabilidade do efeito de parcela.

Os componentes de variância σ_f^2, σ_p^2, σ_δ^2 e σ_{pop}^2 são: variância entre famílias, entre parcelas, dentro de parcelas e entre populações, respectivamente. As quantidades n, b e p referem-se aos números de indivíduos por parcela, número de blocos e número de famílias, respectivamente.

A estimação de σ_a^2 usando dados apenas da geração F_3 implica assumir 0,25 σ_d^2 tendendo a zero na variação entre progênies. Entretanto, mesmo sem esta suposição, a presença desta pequena fração da variância de dominância (σ_d^2) não deverá afetar o *ranking* pelo BLUP, pois tal variância estará incluída (0.125 σ_d^2) também no numerador do peso (b_1) dado ao componente dentro de linhagem, ao se obter 0,50 da variância genética entre linhagens no numerador de b_1. Este índice ou BLUP é adequado também para a seleção envolvendo progênies S1 de espécies alógamas.

Em plantas autógamas geralmente são avaliadas simultaneamente p linhagens pertencentes a várias populações (pop) segregantes. Neste caso, o BLUP para seleção pelos efeitos genéticos aditivos está associado ao seguinte índice multi-efeitos:

$$I_2 = b_1\delta_{ijkl} + b_2 g_i + b_3\, c_{ijl} + b_4 pop_l$$

$$= b_1\left(Y_{ijkl} - \overline{Y}_{ij}..\right) + b_2\left(\overline{Y}_{i\cdot\cdot\cdot} - \overline{Y}\ldots.\right) + b_3\left(\overline{Y}_{ij}.. - \overline{Y}_{i}... - \overline{Y}_{.j}.. + \overline{Y}\ldots.\right) + b_4\left(\overline{Y}_{\cdot\cdot l} - \overline{Y}\ldots.\right),$$ em que

$b_4 = \dfrac{\sigma^2_{pop}}{\sigma^2_{pop} + \sigma^2_p / p + \sigma^2_c / bp + \sigma^2_\delta / nbp}$ é a herdabilidade do efeito da média da população.

No caso em que as linhagens F_3 são semeadas em linha e não há repetição, o índice ótimo dentro de população equivale a $I_3 = b_5\left(Y_{ij} - \overline{Y}_{i\cdot}\right) + b_6\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)$, em que b_5 = b_1 e $b_6 = \dfrac{\left[2n+1)/(2n)\right]\sigma_a^2}{\sigma_p^2 + \sigma_\delta^2 / n}$. É importante relatar que neste caso a seleção não é puramente genética, pois o experimento não teve repetição, fato que prejudica também a casualização.

7.2 Método BLUPIS na Seleção em Cana-de-Açúcar e Plantas Forrageiras

Em espécies de reprodução vegetativa (como a cana-de-açúcar e espécies forrageiras), o procedimento ideal de seleção de indivíduos para clonagem na fase inicial do melhoramento é o BLUP individual considerando simultaneamente as informações do indivíduo, da família, do delineamento experimental e do parentesco entre famílias e genitores. Entretanto, a informação do indivíduo geralmente não é obtida por ocasião da avaliação das famílias, as quais são avaliadas por meio de colheita total das parcelas.

O valor genotípico verdadeiro, intrínseco ou paramétrico desses indivíduos não avaliados, considerando o indivíduo i da família j, é dado por $u + g_{ij} = u + g_j + g_{i/j}$, em que u é a média geral, g_{ij} é o efeito genotípico do indivíduo ij, g_j é o efeito genotípico da família j e $g_{i/j}$ é o desvio genotípico do indivíduo i dentro da família j. Esta expressão pode ser rescrita como $u + g_{ij} = u + g_j + h^2_{gd}(y_{ij} - g_j) = u + g_j(1 - h^2_{gd}) + h^2_{gd} y_{ij}$, em que y_{ij} é a observação fenotípica do indivíduo ij e h^2_{gd} é a herdabilidade genotípica dentro de família de irmãos germanos, cujo numerador é dado por $(1/2)\sigma^2_a + (1/4)\sigma^2_d$. O BLUP de u + g_{ij} é dado por $\hat{u} + \hat{g}_{ij} = \hat{u} + \hat{g}_j + h^2_{gd}(y_{ij} - \hat{g}_j) = \hat{u} + \hat{g}_j(1 - h^2_{gd}) + h^2_{gd} y_{ij}$ em que $\hat{g}_j$ é o BLUP para famílias de irmãos germanos, obtido após consideração do parentesco entre as famílias e entre os genitores envolvidos na avaliação genética. Mas como y_{ij} não foi observado, tal BLUP não pode ser calculado explicitamente. Mas a comparação entre os BLUP´s de dois indivíduos distintos ij e lk, pertencentes às famílias j e k, pode ser realizada. No caso, o indivíduo da família j será superior ao indivíduo da família k se $\hat{u} + \hat{g}_j(1 - h^2_{gd}) + h^2_{gd} y_{ij} > \hat{u} + \hat{g}_k(1 - h^2_{gd}) + h^2_{gd} y_{lk}$. Percebe-se que as quantidades $h^2_{gd} y_{ij}$ e $h^2_{gd} y_{lk}$, ou seja, as frações de y ditadas pela herdabilidade dentro de família, independem completamente dos valores genotípicos $\hat{g}_j$ e $\hat{g}_k$ das famílias e são completamente aleatórias pois são efeitos da segregação mendeliana. Assim sendo, $h^2_{gd} y_{ij}$ e $h^2_{gd} y_{lk}$ tem igual esperança matemática $h^2_{gd} y$. Assim, em média ou esperança matemática, o indivíduo da família j será superior se $\hat{u} + \hat{g}_j(1 - h^2_{gd}) + h^2_{gd} y > \hat{u} + \hat{g}_k(1 - h^2_{gd}) + h^2_{gd} y$, ou seja, se $\hat{g}_j > \hat{g}_k(1 - h^2_{gd})/(1 - h^2_{gd}) + (h^2_{gd} y - h^2_{gd} y + \hat{u} - \hat{u})/(1 - h^2_{gd})$, portanto se $\hat{g}_j > \hat{g}_k$, ou seja, se $\hat{g}_j - \hat{g}_k > 0$ ou ainda

se $\hat{g}_j / \hat{g}_k > 1$. Dessa forma, $\hat{g}_j / \hat{g}_k$ indica a taxa média de indivíduos superiores na família j em relação aos indivíduos da família k.

Se $\hat{g}_j / \hat{g}_k$ = 1.2 e são selecionados 40 indivíduos por família k, deverão ser selecionados 48 indivíduos da família j para que o pior indivíduo selecionado da família j tenha o mesmo nível do pior indivíduo selecionado da família k. E, no caso, estes 88 indivíduos deverão coincidir aproximadamente com os 88 melhores indivíduos que teriam sido selecionados pelo BLUP aplicado na seleção de indivíduos pertencentes a estas duas famílias. Em resumo, a determinação do número de indivíduos a serem selecionados em cada família, usando a relação entre os efeitos genotípicos das famílias de irmãos germanos, simulará bem a seleção pelo procedimento BLUP individual. Por isto tal procedimento é denominado BLUP individual simulado (BLUPIS) e a expressão que determinará de forma dinâmica o número n_k de indivíduos selecionados em cada família k é dado por $n_k = (\hat{g}_k / \hat{g}_j) n_j$ em que $\hat{g}_j$ refere-se ao valor genotípico da melhor família e n_j equivale ao número de indivíduos selecionados na melhor família (Resende, 2004; Resende e Barbosa, 2005; 2006). alternativamente, tal expressão pode ser dada por $n_k = [1 - (\hat{g}_j - \hat{g}_k)/(\hat{g}_j)] n_j = (\hat{g}_k / \hat{g}_j) n_j$. Por esta última expressão verifica-se que n_k depende do tamanho da diferença entre os efeitos genotípicos das duas famílias como proporção do efeito genotípico da melhor família. O BLUPIS é um melhoramento da seleção seqüencial em cana-de-açúcar, a qual é amplamente adotada na Austrália, Brasil, Estados Unidos e Argentina. A determinação de n_j envolve o conceito de tamanho efetivo populacional (Ne) e pode ser tomado como 50, o qual representa 98 % do tamanho efetivo máximo de uma família de irmãos germanos. O Ne para famílias de irmãos germanos é dado por *Ne = (2n)/(n+1)*, conforme Vencovsky (1978).

O método elimina automaticamente as famílias com efeito genotípico negativo, ou seja, aquelas abaixo da média geral do experimento. Isto é razoável quando se considera a baixíssima probabilidade de se obter um clone superior nestas famílias. Esta abordagem é também adequada a outras espécies de reprodução vegetativa como a braquiária, o *Panicum*, o capim elefante, a mandioca. Também é adequado a todas as espécies autógamas anuais e perenes (café arábica), as quais são avaliadas em nível de totais de parcelas. Nessas espécies, o número de linhagens irmãs a serem avançadas, ou seja, o número de indivíduos a serem selecionados e avançados dentro de

cada linhagem, pode ser determinado pelo BLUPIS. No caso de famílias F_3 ou S_1, o n_j pode ser tomado como 20, o qual representa 98 % do Ne máximo de uma família F_3 ou S_1. O Ne para famílias S_1 é dado por *Ne = n/(n + 0,5)*.

O BLUPIS fornece também a informação de qual parcela ou em qual repetição se encontram os genótipos superiores de cada família ou acesso. Para isso é usada a predição dos efeitos genotípicos de cada parcela de cada família, que é dada por $\hat{g}_{parc_r} = (y - X\hat{r} - W\hat{b} - Z\hat{g}) * h^2_{parc} = \mathrm{Re}\, s_r * h^2_{parc}$, considerando a avaliação experimental no delineamento de blocos incompletos, em que r, b e g referem-se aos efeitos de repetições, blocos e genotípicos de famílias de irmãos germanos, respectivamente. Res_r é o resíduo da parcela r. A herdabilidade dos efeitos de parcela (h^2_{parc}) é dada por $h^2_{parc} = (\sigma^2_{gd} / n)/(\sigma^2_e)$, em que σ^2_{gd} é a variância genotípica dentro de famílias de irmãos germanos, n é o número de plantas em cada parcela e σ^2_e é a variância residual. No caso em que se usa uma observação média por parcela, σ^2_e contempla (σ^2_{gd} / n) e a variância ambiental. No caso em que se usam observações individuais dentro de parcelas, os efeitos genotípicos de parcela são dados por $\hat{g}_{parc_r} = \hat{p}arc_r * [(\sigma^2_{gd} / n)/(\sigma^2_{parc})]$ em que σ^2_{parc} é a própria variância ambiental entre parcelas ajustada pelo modelo e $\hat{p}arc_r$ é o próprio efeito da parcela r ajustado pelo modelo.

De posse da predição dos efeitos genotípicos de cada parcela, o número de indivíduos a ser selecionado de cada parcela r é dado pela proporção $(u + \hat{g}_i + \hat{g}_{parc_r})/[\sum(u + \hat{g}_i + \hat{g}_{parc_r})]$ multiplicada pelo número total de indivíduos a ser selecionado por família, fornecido pelo BLUPIS. A quantidade $\sum(u + \hat{g}_i + \hat{g}_{parc_r})$ refere-se ao somatório dos valores genotípicos $(u + \hat{g}_i + \hat{g}_{parc_r})$ de todas as parcelas de uma família i.

7.3 Método BLUPIS - BIEFEITOS na Seleção em Cana-de-Açúcar e Plantas Forrageiras

Um procedimento melhorado do BLUPIS é o método Blupis – Biefeitos, que combina o BLUP de família ($\hat{g}_{fam} = \hat{g}_i$) com o BLUP de parcela ($\hat{g}_{parc_r}$), visando simular o BLUP individual de clones potenciais pertencentes às famílias em avaliação. É, então, dado por $\hat{g}_{(fam+parc)} = \hat{g}_{fam} + \hat{g}_{parc_r}$. Assim, a determinação do número n_{kr} de indivíduos a ser selecionado por combinação família k–parcela deve basear-se na relação $n_{kp} = (\hat{g}_{(fam+parc)kr} / \hat{g}_{(fam+parc)jr})n_{jr}$ em que $\hat{g}_{(fam+parc)j}$ refere-se ao valor genotípico da melhor combinação família-parcela da melhor família no experimento e n_{jr} equivale ao número de indivíduos selecionados nessa melhor combinação família–parcela da melhor família. Esse número n_{jr} é determinado pela relação $[(u + \hat{g}_j + \hat{g}_{parc_r}) / [\sum (u + \hat{g}_j + \hat{g}_{parc_r})]] * n_j$ para a melhor família, sendo n_j da ordem de 50, conforme relatado no tópico anterior. A seleção deve incluir apenas parcelas com $\hat{g}_{(fam+parc)}$ positivo. O Blupis – Biefeitos é tanto mais vantajoso que o Blupis quanto menor for o número de plantas por parcela usado na experimentação.

7.4 Método BLUP–VEG Individual para Espécies de Propagação Vegetativa

Esse método é aplicável na seleção individual em todas as espécies de propagação assexuada, por via vegetativa ou apomixia, englobando, portanto, espécies florestais, frutíferas, forrageiras e algumas olerícolas e ornamentais. Aplica-se também em espécies autógamas perenes como o café arábica e o pessegueiro. Pode ser aplicado com famílias de meios irmãos, irmãos germanos, irmãos germanos sob cruzamentos dialélicos e fatoriais, testes de acessos ou procedências.

O método BLUP-VEG reúne princípios do BLUP-HET, do BLUP individual e do BLUP da segregação ou amostragem mendeliana (BLUP-SAM). O BLUP-VEG é individual e dado por BLUP-VEG = BLUP-HET de família + BLUP-SAM dentro de família. O método utiliza também o BLUP da variância genética residual dentro de famílias, ou seja, emprega não apenas BLUP de componentes de médias, mas também BLUP de componentes de variância.

Um preditor BLUP de efeitos genotípicos individuais, com acurácia máxima, deve contemplar as diferentes quantidades de variação genética dentro das diferentes famílias, ou seja, a heterogeneidade de variância genética dentro das famílias. Essa heterogeneidade é decorrente dos diferentes níveis de segregação ou amostragem mendeliana de genes dentro de famílias. O BLUP individual considerando essa heterogeneidade é composto de duas partes: BLUP-HET dos efeitos genotípicos de família + BLUP-HET dos efeitos de indivíduo dentro de família. Assim, a heterogeneidade é considerada em duas etapas: entre famílias, contemplando a heterogeneidade de variância fenotípica residual; dentro de famílias, contemplando separadamente a heterogeneidade de variância genética e fenotípica residuais.

A separação da variância genética residual da variância residual total dentro de cada família é necessária para obtenção do componente BLUP-SAM do BLUP-VEG. O BLUP-SAM usa uma herdabilidade individual no sentido amplo dentro de família, específica para cada família. Os denominadores dessas herdabilidades são as próprias variâncias fenotípicas residuais dentro de cada família i ($\hat{\sigma}^2_{fd_i}$) e os numeradores devem ser determinados por $\hat{\sigma}^2_{gd_i} = \hat{\sigma}^2_{fd_i} - \hat{\sigma}^2_{ed_i}$. O componente ambiental $\hat{\sigma}^2_{ed_i}$ pode ser assumido como constante para todas as famílias, ou seja, pode-se assumir que a variação ambiental é a mesma ($\hat{\sigma}^2_{ed}$) para todos os indivíduos. Dessa forma, componente $\hat{\sigma}^2_{ed}$ pode ser estimado a partir da avaliação de alguns clones (pelo menos cinco e não pertencentes às famílias em avaliação) em conjunto com as progênies e $\hat{\sigma}^2_{ed}$ será estimado como a média das variâncias residuais dentro desses clones, em um modelo que analisa simultaneamente as progênies e clones. Uma aproximação para obter $\hat{\sigma}^2_{ed}$ é tomá-la como a menor variância fenotípica residual dentro de famílias, ou seja, aquela associada à família menos variável. Isso não é exato, mas é um procedimento superior àquele em que se ignora a heterogeneidade de variância.

Uma terceira abordagem é a obtenção direta da variância genotípica dentro de famílias ($\hat{\sigma}^2_{gd_i}$) como um BLUP de componente de variância, a partir da análise de variância das próprias variâncias residuais dentro de famílias. Esse procedimento considera o controle genético das variâncias residuais ou segregação mendeliana. Nesse caso, o BLUP-VEG é dado por BLUP-VEG = BLUP-HET de família + BLUP-SAM dentro de família = BLUP-HET de família + [(BLUP de ($\sigma^2_{gd_i}$)) / $\hat{\sigma}^2_{fd_i}$] *

$\mathrm{Re}\, s_{ij}$, em que $\mathrm{Re}\, s_{ij}$ é o resíduo associado ao indivíduo j da família i. Preditores BLUP-VEG individual contemplando simultaneamente a heterogeneidade de variâncias, com herdabilidades entre e dentro de famílias específicas para cada família, podem ser obtidos pelo *software* Selegen-Reml/Blup.

O BLUP de $\sigma^2_{gd_i}$ é dado por $BLUP(\sigma^2_{gd_i}) = \sigma^2_{gd_m} + [Var_g(\sigma^2_{fd_i}) / Var_f(\sigma^2_{fd_i})] * (\sigma^2_{fd_i} - \sigma^2_{fd_m})$, em que $\sigma^2_{gd_m}$ e $\sigma^2_{fd_m}$ são as variâncias genética e fenotípica dentro de progênies, médias para todas as progênies. O termo $Var_g(\sigma^2_{fd_i})$ refere-se à variância genética entre progênies das variâncias fenotípicas dentro de progênies. Por sua vez, o termo $Var_f(\sigma^2_{fd_i})$ refere-se à variância fenotípica entre progênies das variâncias fenotípicas dentro de progênies.

É importante relatar que, quando os dados associados ao caráter em questão tem distribuição normal, as variâncias residuais dentro de progênies tem distribuição qui-quadrado. Assim, com homogeneidade de variâncias genéticas, tem-se $Var_f(\sigma^2_{fd_i}) = 2(\sigma^2_{fd_m})^2 / gl_i$, e com heterogeneidade de variâncias genéticas, tem-se $Var_f(\sigma^2_{fd_i}) = 2(\sigma^2_{fd_m})^2 / gl_i + Var_g(\sigma^2_{fd_i})$, em que gl_i é o número de graus de liberdade associado à estimação de cada variância residual dentro de progênie. Isto porque a variância de amostragem de uma variável com distribuição qui-quadrado equivale a $2u^2 / gl_i$ em que u é a média geral. A obtenção de $Var_f(\sigma^2_{fd_i})$ baseia-se na variância das próprias $\sigma^2_{fd_i}$. Como $\sigma^2_{fd_m}$ é obtida como a própria média das $\sigma^2_{fd_i}$, $Var_g(\sigma^2_{fd_i})$, pode ser obtida como $Var_g(\sigma^2_{fd_i}) = Var_f(\sigma^2_{fd_i}) - [2(\sigma^2_{fd_m})^2 / gl_i]$.

Adotando-se uma transformação logarítmica para eliminar alguma correlação que possa existir entre média e variância, com homogeneidade de variâncias genéticas, tem-se $Var_f[\ln(\sigma^2_{fd_i})] = 2 / gl_i$, e com heterogeneidade de variâncias genéticas, tem-se $Var_f[\ln(\sigma^2_{fd_i})] = 2 / gl_i + Var_g(\sigma^2_{fd_i}) / (\sigma^2_{fd_m})^2$. Isto porque $\ln(\sigma^2_{fd_i})$ tem distribuição assintótica normal com variância $2 / gl_i$.

Para obtenção de $BLUP(\sigma^2_{gd_i})$ resta, então, obter uma estimativa de $\sigma^2_{gd_m}$. Em testes de progênies advindas de cruzamentos dialélicos e fatoriais, $\sigma^2_{gd_m}$ equivale a $(1/2)\sigma^2_a + (3/4)\sigma^2_d$, em que

σ_a^2 e σ_d^2 referem-se à variância genética aditiva e de dominância estimadas a partir da análise do dialelo ou do fatorial. Para o caso de famílias de meios irmãos e de irmãos germanos, $\sigma^2_{gd_m}$ pode ser obtida multiplicando-se $\sigma^2_{fd_m}$ por $[(3/4)h_a^2 + h_g^2]/(1-0{,}25h_a^2)$ e $[(1/2)h_a^2 + (3/4)(h_g^2 - h_a^2)]/[(1-0{,}5h_a^2 - 0{,}25(h_g^2 - h_a^2)]$, respectivamente, em que h_a^2 e h_g^2 são valores conhecidos da herdabilidade individual no sentido restrito e amplo, respectivamente.

Uma alternativa mais simples, porém aproximada, é a obtenção direta de $BLUP(\sigma^2_{gd_i})$ como a variância dos efeitos genotípicos de indivíduos dentro de cada família, obtidos sob um modelo assumindo homogeneidade de variância. E então aplicar a expressão BLUP-VEG = BLUP-HET de família + BLUP-SAM dentro de família = BLUP-HET de família + [(BLUP de ($\sigma^2_{gd_i}$)) / $\hat{\sigma}^2_{fd_i}$] * $\mathrm{Re}\, s_{ij}$.

A variabilidade genotípica diferenciada dentro de famílias ocorre devida aos diferentes graus de homozigose dos genitores. A variação genética aditiva dentro de famílias de irmãos completos é causada por amostragem mendeliana nos locos segregantes dos genitores. Metade dela é causada pelo genitor masculino (variabilidade entre gametas em torno do valor médio do genitor masculino) e a outra metade pelo genitor feminino. A variabilidade causada por um genitor não endógamo é, em média, $(1/4)\sigma_a^2$. Mas se o genitor tem coeficiente de endogamia F, essa variabilidade equivale a $[(1-F)/4]\sigma_a^2$. Isto é devido ao fato de que uma fração F dos locos é ocupada por alelos idênticos por descendência e, portanto, não contribuem para a variabilidade gerada pela segregação. Assim, a variabilidade genética aditiva dentro de famílias de irmãos completos é dada por $\sigma^2_{ad} = [(1-F_m)/4]\sigma_a^2 + [(1-F_f)/4]\sigma_a^2$, em que F_m e F_f são os coeficientes de endogamia dos genitores masculinos e femininos, respectivamente. Assim, a variabilidade genética dentro das famílias difere entre famílias, devido aos diferentes coeficientes de endogamia dos genitores envolvidos em cada uma.

Sob o modelo genético infinitesimal e com genitores não endógamos, a heterogeneidade de variância genética residual dentro de famílias é devido à segregação de genes de grandes efeitos. Assim, a heterogeneidade de variância residual dentro de famílias pode ser usada para inferência sobre o número de locos de características quantitativas que estão segregando. Rowe et al. (2006)

relatam que o coeficiente de variação genética entre famílias de meios irmãos obtidas sob cruzamentos em pares simples (dialélico, fatorial, hierárquico), associado à heterogeneidade de variância residual dentro de famílias, $CV_{g(\sigma^2_{fd_i})} = [Var_g(\sigma^2_{fd_i})]^{1/2} / (\sigma^2_{fd_m})$, pode ser usado na inferência sobre o número de QTL´s segregando. Para freqüências alélicas intermediárias e genes de efeitos iguais, a relação $CV_{g(\sigma^2_{fd_i})} / h^2_a = [1/(4n^{1/2})]$ indica o número n de locos, de efeitos aditivos (caráter com controle genético aditivo), segregantes. E a expressão $CV_{g(\sigma^2_{fd_i})} / h^2_a = [1/(2n^{1/2})]$ indica o número n de locos, de efeitos dominantes (caráter com dominância completa), segregantes.

Apresenta-se a seguir o número n de locos dominantes controlando um caráter, de acordo com sua herdabilidade individual no sentido restrito (h^2_a), coeficiente $CV_{g(\sigma^2_{fd_i})}$ e relação $CV_{g(\sigma^2_{fd_i})} / h^2_a$.

h^2_a	$CV_{g(\sigma^2_{fd_i})}$	$CV_{g(\sigma^2_{fd_i})} / h^2_a$	n	h^2_a	$CV_{g(\sigma^2_{fd_i})}$	$CV_{g(\sigma^2_{fd_i})} / h^2_a$	N	h^2_a	$CV_{g(\sigma^2_{fd_i})}$	$CV_{g(\sigma^2_{fd_i})} / h^2_a$	N
0.1	**0.01**	**0.10**	**25.00**	**0.3**	**0.01**	**0.03**	**225.00**	**0.5**	**0.01**	**0.02**	**625.00**
0.1	0.05	0.50	1.00	0.3	0.05	0.17	9.00	0.5	0.10	0.20	6.25
0.1	0.10	1.00	0.25	0.3	0.10	0.33	2.25	0.5	0.15	0.30	2.78
0.1	0.15	1.50	0.11	0.3	0.15	0.50	1.00	0.5	0.20	0.40	1.56
0.1	0.20	2.00	0.06	0.3	0.20	0.67	0.56	0.5	0.25	0.50	1.00
0.1	0.25	2.50	0.04	0.3	0.25	0.83	0.36	0.5	0.30	0.60	0.69
0.1	0.30	3.00	0.03	0.3	0.30	1.00	0.25	0.5	0.40	0.80	0.39
0.1	0.40	4.00	0.02	0.3	0.40	1.33	0.14	0.5	0.50	1.00	0.25
0.1	0.50	5.00	0.01	0.3	0.50	1.67	0.09	0.6	0.01	0.02	900.00
0.2	**0.01**	**0.05**	**100.00**	**0.4**	**0.01**	**0.03**	**400.00**	**0.6**	**0.10**	**0.17**	**9.00**
0.2	0.05	0.25	4.00	0.4	0.05	0.13	16.00	0.6	0.15	0.25	4.00
0.2	0.10	0.50	1.00	0.4	0.10	0.25	4.00	0.6	0.20	0.33	2.25
0.2	0.15	0.75	0.44	0.4	0.15	0.38	1.78	0.6	0.25	0.42	1.44
0.2	0.20	1.00	0.25	0.4	0.20	0.50	1.00	0.6	0.30	0.50	1.00
0.2	0.25	1.25	0.16	0.4	0.25	0.63	0.64	0.6	0.40	0.67	0.56
0.2	0.30	1.50	0.11	0.4	0.30	0.75	0.44	0.6	0.50	0.83	0.36
0.2	0.40	2.00	0.06	0.4	0.40	1.00	0.25	0.5	0.01	0.02	625.00

Verificam-se os seguintes fatos: (i) para um $CV_{g(\sigma^2_{fd_i})}$ fixo, quanto maior a herdabilidade, maior o número de locos controlando o caráter; (ii) para uma herdabilidade fixa, quanto menor $CV_{g(\sigma^2_{fd_i})}$, maior o número de locos controlando o caráter e segregando na população; (iii) para uma herdabilidade fixa, quando o $CV_{g(\sigma^2_{fd_i})}$ tende a zero (ausência de heterogeneidade de variância genética

dentro de famílias), o número de locos controlando o caráter tende a infinito e o modelo genético infinitesimal se aplica perfeitamente ao caráter quantitativo, ou seja, não há genes de grandes efeitos; (iv) para uma herdabilidade fixa, quanto maior $CV_{g(\sigma^2_{fdi})}$, menor o número de locos de grande efeito controlando o caráter; (v) para uma herdabilidade fixa, quando o valor de $CV_{g(\sigma^2_{fdi})}$ atinge 50 % do valor da herdabilidade, um gene de grande efeito está segregando na população; (vi) para uma herdabilidade fixa, quando o valor de $CV_{g(\sigma^2_{fdi})}$ ultrapassa 50 % do valor da herdabilidade, nenhum gene de grande efeito está segregando na população e deve-se assumir que a variância ambiental também está sob controle genético.

Para caracteres com controle genético puramente aditivo, tendências similares são observadas. Porém, o número de locos igual a 1 é obtido quando o valor de $CV_{g(\sigma^2_{fdi})}$ atinge 25 % do valor da herdabilidade. Porcentagens acima de 25 % já indicam que a variância ambiental tem controle genético.

8 RESTRIÇÕES DE SOMATÓRIO NULO ASSOCIADAS AO AJUSTE DE EFEITOS ALEATÓRIOS E FIXOS

No uso de modelos mistos é preciso enxergar como os efeitos estão sendo ajustados e se existem graus de liberdade suficientes para ajustá-los. Não basta escolher as colunas classificatórias dos arquivos de dados como efeitos fixos e aleatórios e efetuar uma análise em um programa computacional. Em algumas situações, a própria forma de codificar as colunas determina quais modelos de fato estão sendo ajustados. Inclusive modelos estatísticos diferentes podem ser ajustados para diferentes materiais genéticos em um mesmo arquivo de dados. Maiores detalhes podem ser vistos em Resende (2007).

A seguir são apresentadas algumas restrições associadas às soluções BLUP, que podem ser úteis na tentativa de se enxergar melhor os modelos que estão sendo ajustados. Essas restrições foram apresentadas por Searle (1997 a e b).

Tais restrições referem-se ao fato de que certas somas de alguns elementos BLUP somam zero. Este fato não decorre de restrições impostas nos parâmetros do modelo tal como ocorre no contexto dos modelos de efeitos fixos, mas decorre da própria forma dos preditores BLUP, e a ocorrência de dados perdidos não afeta isso. Tais restrições são:

a) Fatores aleatórios: para cada fator de efeitos aleatórios, a soma dos BLUP´s de seus efeitos equivale a zero;

b) Fator referente à interação entre um fator de efeitos fixos e outro fator de efeitos aleatórios: para cada nível do fator de efeitos fixos, a soma dos BLUP´s referente à referida interação equivale a zero.

Um exemplo da restrição (b) refere-se ao fator parcela γ_{ij} (interação entre progênie como fator de efeitos aleatórios j e bloco como fator de efeitos fixos i, por exemplo) em um experimento no delineamento em blocos ao acaso. Nesse caso, a soma dos efeitos de parcela equivale a zero dentro de cada bloco. Matematicamente, tem-se $\sum_j \gamma_{ij} = 0$ para cada bloco i. Logicamente, o somatório total para todos os blocos também equivale a zero. O mesmo é válido para o caso da interação genótipos x locais, com progênies de efeitos aleatórios e locais de efeitos fixos.

A restrição (b) aplica-se apenas no caso de interação entre fator de efeitos fixos e fator de efeitos aleatórios. Não se aplica no caso de interação entre dois fatores aleatórios. Nesse caso, a restrição (a) se aplica.

No caso de interação envolvendo mais que dois fatores, sendo um deles de efeitos fixos, a restrição (b) também se aplica. Por exemplo, no caso da interação genótipos x locais x colheitas em plantas perenes. Considerando locais como de efeitos fixos e os demais como aleatórios, as interações genótipos x colheitas somam zero dentro de cada local.

A restrição (b) aplica-se também ao caso de efeitos aleatórios hierárquicos dentro de efeitos fixos. Por exemplo, no caso de progênies tomados como de efeitos aleatórios dentro de populações tomadas como de efeitos fixos. Nesse caso, se existe uma só progênie para representar a população, o efeito da progênie será ajustado como zero e toda a informação disponível será sugada pelo efeito fixo da população.

CAPÍTULO 4

REML: ASPECTOS MATEMÁTICOS, ESTATÍSTICOS E COMPUTACIONAIS

Inferências estatísticas fidedignas dependem da qualidade dos dados, de uma modelagem plausível e da adoção de procedimentos adequados de estimação e predição. Nos capítulos anteriores abordou-se a questão da qualidade dos dados experimentais e da modelagem adequada. Neste capítulo, ênfase é dada a um procedimento ótimo de estimação e predição. Esse refere-se ao método de máxima verossimilhança residual (REML) desenvolvido por Patterson e Thompson (1971) e Thompson (1973; 1980), a partir de esforços na tentativa de obtenção de melhores estimadores de componentes de variância associados a dados não ortogonais e desbalanceados (Thompson, 1969). O REML é um método fisheriano de modelagem e inferência mas pode ser derivado sob o enfoque bayesiano, fato que caracteriza a sua generalidade como procedimento ótimo. Fundamentos e detalhes dos procedimentos de máxima verossimilhança são apresentados em livros específicos sobre o assunto tais quais Edwards (1992), Severini (2001), Sorensen e Gianola (2002), Pawitan (2001), Foulley (2003) e Lee et al. (2007). A abordagem apresentada aqui baseia-se, principalmente, nesses livros e também nos livros de Lynch e Walsh (1998) e Resende (2002).

Em geral, o procedimento REML é aplicado em conjunto com a técnica BLUP apresentada formalmente por Henderson (1973; 1975; 1984) e Thompson (1976; 1977; 1979). Esses dois autores contribuíram muito para a evolução da estatística na área de modelos lineares e não lineares mistos. As maiores autoridades nessa área foram alunos desses dois cientistas tais quais Searle, Harville, Quaas, Schaeffer e Foulley, que foram alunos de Henderson e Karin Meyer e Mrode que foram alunos e colaboradores de Thompson. Searle foi professor de Henderson em álgebra de matrizes na Nova Zelândia e depois foi aluno de Henderson nos EUA. Thompson trabalhou muito em conjunto com Gilmour e Cullis da Austrália. Esses três e Karin Meyer estão dentre os maiores especialistas em REML. Outros pesquisadores importantes na área de modelos mistos e métodos de seleção aplicados ao melhoramento genético são Van Vleck (Van Vleck et al. 1987; Van Vleck, 1993), Piepho (vários trabalhos referenciados nesse documento), Misztal (Misztal, 1999) e Gallais (Gallais, 1989).

Gianola (2006) relata que os principais cientistas que contribuíram para a criação e desenvolvimento de métodos estatísticos para a análise de dados foram Pearson, Fisher, Lush, Henderson, Searle e Thompson. A esses nomes pode ser acrescentado o próprio Gianola como introdutor do paradigma bayesiano no melhoramento genético. Vários importantes trabalhos de todos esses autores são citados no presente livro.

1 INFERÊNCIA VEROSSIMILHANÇA, INFERÊNCIA FREQÜENTISTA E INFERÊNCIA BAYESIANA

A inferência estatística tem evoluído muito nas últimas décadas, principalmente em função do aumento dos recursos computacionais. Desde o início do desenvolvimento da estatística até recentemente, houve grande supremacia da **inferência freqüentista**, que, em geral, tem sido creditada como fisheriana e se apóia, principalmente, na consistência assintótica através da qual um modelo estimado converge para o verdadeiro modelo. A propriedade de incorporar a informação a priori na análise, sob o enfoque bayesiano, sempre gerou muita controvérsia, em determinada fase do desenvolvimento da estatística (Efron, 1986; Lindley, 1978; Smith, 1984). Atualmente, este tipo de discussão parece não ser mais relevante, com as diferenças entre as duas abordagens,

frequentista e bayesiana, bem compreendidas e as virtudes de cada uma delas utilizadas quando mais conveniente (Gamerman, 1996).

A inferência freqüentista é fortemente baseada na estimação por quadrados mínimos. Atualmente, uma das marcantes características da inferência estatística moderna é o papel central da função de verossimilhança, a qual foi introduzida por Fisher (1922). Segundo Lindsey (1999), poucos estatísticos modernos percebem que Fisher nunca foi um freqüentista para inferência (embora ele tenha geralmente usado uma interpretação freqüentista de probabilidade), sendo muito mais próximo aos bayesianos, tais quais Jeffreys, do que da escola de Neyman-Pearson (os verdadeiros freqüentistas, que introduziram a teoria dos testes de hipóteses entre 1936 e 1938). Segundo tal autor, a inferência estatística no sentido fisheriano enfatiza a obtenção do máximo de informação dado o conjunto de observações, sem incorporação de conhecimento e informação a priori, exceto aquelas necessárias para a construção de modelos (o princípio da verossimilhança baseia-se na suposição de que o modelo ou a função são verdadeiros e que somente os valores corretos dos parâmetros necessitam ser determinados). Portanto, a inferência Fisheriana refere-se, essencialmente, à inferência verossimilhança.

O uso da função de verossimilhança como um meio direto de fazer inferência, conforme sugerido por Fisher, é um enfoque recente em estatística (Edwards, 1972; Bardorff-Nielsen, 1976; Gomes, 1981; Pereira, 1997; Lindsey, 1999), caracterizando a inferência verossimilhança.

Pereira (1997) caracteriza muito bem os pontos fortes da inferência verossimilhança, tais quais: (i) a função de verossimilhança fornece uma medida exata da incerteza, resumindo toda a informação que os dados fornecem sobre um parâmetro desconhecido θ; (ii) a verossimilhança é uma medida relativa, de forma que não é possível obter uma medida de plausibilidade absoluta e, por isto, somente as razões de verossimilhança são relevantes; (iii) a função de verossimilhança relativa, dada por $R(\theta;\ y) = \dfrac{\theta^{y}(1-\theta)^{n-y}n^{n}}{y^{n}(n-y)^{n-y}}$ pode ser utilizada para obtenção de níveis de significância aproximados, através dos testes da razão de verossimilhança (veja descrição dos elementos da equação no tópico seguinte); (iv) existe uma relação entre a função de verossimilhança e os métodos convencionais de teoria de grandes amostras, ou seja, além de fornecer uma medida exata de plausibilidade, a função de verossimilhança dá uma indicação de

aplicabilidade de resultados assintóticos em casos particulares; (v) os métodos de verossimilhança são exatos tanto para pequenas quanto para grandes amostras, não dependendo de propriedades assintóticas das várias distribuições amostrais.

Assim, em inferência bayesiana, certos métodos que assumem distribuições a priori não informativas são essencialmente de inferência verossimilhança, os quais mantêm a propriedade de conduzir a análise exata de amostra de tamanho finito.

2 FUNÇÃO DE VEROSSIMILHANÇA

A função de verossimilhança formaliza a contribuição dos dados amostrais para o conhecimento sobre θ, visto que conecta a distribuição a priori à distribuição a posteriori.

A função de verossimilhança [$\ell(\theta; y)$] de θ é a função que associa a cada θ, o valor f(y|θ). Dessa forma $\theta \rightarrow \ell(\theta; y) = f(y \mid \theta)$. A função $\ell(\theta; y)$ associa (para um valor fixo de y) a probabilidade de ser observado y a cada valor de θ. Assim, quanto maior o valor de ℓ maiores são as probabilidades atribuídas pelo particular valor de θ considerado, ao valor fixado de y. Ao fixar um valor de y e variar os valores de θ, observa-se a plausibilidade ou verossimilhança de cada um dos valores de θ. É interessante observar que: $\int_R f(y \mid \theta)\ d\ y = 1\ mas \int_\theta \ell(\theta; y)\ d\ \theta = K \neq 1$, ou seja, a integral da função densidade de probabilidade equivale a 1, mas a função de verossimilhança não integra 1 (Gamerman e Migon, 1993).

Os termos probabilidade e verossimilhança são conceitos diferentes. No cômputo da verossimilhança, fixa-se a amostra ou conjunto de dados (y) e varia-se o parâmetro θ, procurando encontrar o parâmetro verossímil ou plausível com o conjunto de dados. Por outro lado, no cálculo de uma probabilidade, utiliza-se uma distribuição com parâmetro θ conhecido e calcula-se a probabilidade de observar um determinado valor $y = y_0$. Em outras palavras, o problema da probabilidade é predizer a chance de ocorrer y sabendo θ e o problema da verossimilhança é fazer afirmações sobre θ com base no valor observado y.

No cálculo de probabilidades fixa-se θ (conhecido) e varia-se y, ao passo que no cálculo da verossimilhança fixa-se y e varia-se θ.

Como exemplo, considere a distribuição y ~ Binomial (2, θ), em que:

2 = n : número de repetições do experimento.

θ = parâmetro da binomial = probabilidade de sucesso.

y = 0, 1, 2 : espaço amostral (y é o número de sucessos).

$\theta \in (0,1)$: espaço paramétrico ou domínio de variação do parâmetro θ.

A função de probabilidade equivale a $p(y/\theta) = \binom{2}{y} \theta^y (1 - \theta)^{2-y}$, a qual equivale à própria função de verossimilhança, $\ell(\theta; y)$, quando se observa $y = y_0$. Considerando os diferentes elementos do espaço amostral, são as seguintes as funções de verossimilhança:

(i) se y = 0, então $\ell(\theta; y = 0) = \binom{2}{0} \theta^0 (1 - \theta)^2 = (1 - \theta)^2$;

(ii) se y = 1, então $\ell(\theta; y = 1) = \binom{2}{1} \theta^1 (1 - \theta)^1 = 2\theta (1 - \theta)$;

(iii) se y = 2, então $\ell(\theta; y = 2) = \binom{2}{2} \theta^2 (1 - \theta)^0 = \theta^2$.

Tendo-se observado y = 0, variando-se θ no intervalo de 0 a 1 na função de verossimilhança $\ell(\theta; y = 0) = (1 - \theta)^2$, verifica-se que o ponto de máximo desta função ocorre quando $\theta = 0$. Por outro lado, tendo-se observado y = 1, variando-se θ no intervalo de 0 a 1 na função $\ell(\theta; y = 1) = 2\theta (1 - \theta)$, verifica-se que o ponto de máximo ocorre quando $\theta = ½$. Finalmente, tendo-se observado y = 2, verifica-se que o máximo de $\ell(\theta; y = 2) = \theta^2$ equivale a 1 (Figura 1).

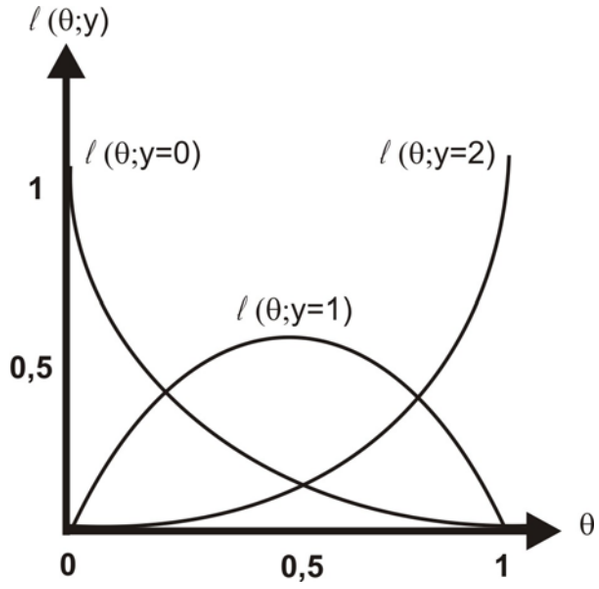


Figura 1. Funções de verossimilhança para diferentes valores observados y = 0, 1, 2.

Em resumo, se for observado:

y = 0 $\Rightarrow \hat{\theta}$ = 0 maximiza ℓ (θ; 0);

y = 1 $\Rightarrow \hat{\theta}$ = ½ maximiza ℓ (θ; 1);

y = 2 $\Rightarrow \hat{\theta}$ = 1 maximiza ℓ (θ; 2).

Genericamente, uma função de verossimilhança pode ser construída conforme descrito a seguir. Sendo, y_1, y_2, ..., y_n uma amostra aleatória independente e identicamente distribuída, com função de probabilidade ou função densidade de probabilidade dada por f ($y|\theta$), em que θ é o parâmetro da função, a função de verossimilhança de $\underset{\sim}{y} = (y_1, y_2, ..., y_n)$ é dada por $\ell(\theta; \underset{\sim}{y}) = f(\underset{\sim}{y}|\theta)$ = f (y_1, y_2, ... $y_n|\theta$), função essa que associa a cada θ, um valor de f($y|\theta$).

A função conjunta $\ell(\theta; \underset{\sim}{y}) = f(\underset{\sim}{y}|\theta)$ = f (y_1, y_2, ... $y_n|\theta$) fatora nas marginais de forma que equivale ao produtório $\ell(\theta; \underset{\sim}{y})$ = f ($y_1|\theta$) ... f ($y_n|\theta$) = $\prod_{i=1}^{n} f(y_i|\theta)$. Como exemplo, considere a

distribuição normal $N(\theta, \sigma^2)$. Neste caso, a função de verossimilhança desta normal equivale a $\ell(\theta; \underset{\sim}{y}) = f(\underset{\sim}{y} \mid \theta) = \prod_{i=1}^{n} f_N(y_i;\ \theta,\ \sigma^2) = \prod_{i=1}^{n} \frac{1}{[2\pi\sigma^2]^{1/2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\theta)^2}{\sigma^2}\right\} \alpha \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\}$, onde $\bar{y}$ é a média aritmética dos y_i.

A função de verossimilhança é a base do Princípio da Verossimilhança, o qual postula que toda a informação contida na amostra ou experimento encontra-se representada nesta função. Existem diferentes verossimilhanças que podem ser empregadas na estimação paramétrica, tais quais verossimilhança incondicional marginal, verossimilhança condicional e verossimilhança parcial. Gianola et al. (1989) apresentam detalhes sobre a estimação envolvendo estes diferentes conceitos.

3 MÁXIMA VEROSSIMILHANÇA (ML)

O método da máxima verossimilhança baseia-se na obtenção do ponto de máximo de uma função de verossimilhança (que é a função densidade de probabilidade conjunta dos pontos amostrais). E este máximo é obtido por derivação da função de verossimilhança (L) em relação ao parâmetro de interesse. Assim, o estimador ML maximiza a verossimilhança do parâmetro dado a função densidade de probabilidade e o conjunto de dados. O ponto de máximo da função de verossimilhança é mais facilmente encontrado quando se toma o logaritmo natural dessa função. Isto porque, com essa transformação, o produtório em $L = \ell(\theta; \underset{\sim}{y})$ relatado no item anterior transforma-se em somatório, fato que torna os cálculos mais tratáveis. No presente texto, as denominações Log e Log_e denotam a mesma coisa, ou seja, o logaritmo natural ou na base e.

O método ML foi desenvolvido por Fisher (1922), mas somente após cerca de 45 anos, Hartley e Rao (1967) apresentaram a especificação matricial de um modelo misto e a derivação de equações ML para várias classes de modelos. Segundo Searle (1991), os trabalhos de Henderson (1953) tiveram grande impacto no desenvolvimento dos métodos de estimação de componentes de variância a partir de dados desbalanceados, estimulando principalmente os trabalhos de Hartley e Rao.

Em situações de dados desbalanceados, os estimadores ML apresentam as seguintes propriedades desejáveis: suficiência, consistência, eficiência e invariância a translação. Essas propriedades são bem descritas em outras obras (Lynch e Walsh, 1998; Resende, 2002). Outra vantagem do ML é a geração de estimativas não negativas dos componentes de variância. Por outro lado, os estimadores ML são viciados em decorrência da perda de graus de liberdade devida à estimação dos efeitos fixos (Shaw, 1987). Para a estimação ML de componentes de variância os efeitos fixos devem ser conhecidos. No entanto, eles não são conhecidos e são substituídos por suas estimativas obtidas por ML. Mas na estimação dos componentes de variância o método ML não considera a perda de graus de liberdade devido à estimação desses efeitos fixos, causando então o vício. Este vício conduz a subestimativas dos parâmetros de variância e portanto podem conduzir a inferências incorretas. As estimativas são viciadas mas possuem menores variâncias de amostragem do que aquelas propiciadas por estimadores não viciados.

Embora viciado, o procedimento ML é computacionalmente mais simples que o método REML (descrito a seguir) e, em determinadas situações, apresenta eficiência satisfatória. O vício pode ser considerável se o número de equações independentes (posto de X, em que X é a matriz de incidência dos efeitos fixos), para os efeitos fixos, for relativamente grande em relação ao número (N) de observações. Quando o posto de X é pequeno em relação a N, os métodos ML e REML conduzem a resultados similares, conforme verificado por Resende et al. (1996b).

4 MÁXIMA VEROSSIMILHANÇA RESIDUAL (REML)

O método REML foi desenvolvido por Patterson e Thompson (1971), que apresentaram uma correção ao ML, eliminando o seu vício. No método REML, somente a porção da verossimilhança que é invariante aos efeitos fixos (especificados no vetor β) é maximizada. Assim, o REML mantém as demais propriedades do ML, é não viciado e permite também a imposição de restrições de não negatividade. Dessa forma, o REML é o procedimento ideal de estimação de componentes de variância com dados desbalanceados. No método REML, os componentes de variância são estimados sem serem afetados pelos efeitos fixos do modelo e os graus de liberdade referentes à estimação dos efeitos fixos são considerados, produzindo estimativas não viciadas (McCulloch e Searle, 2001).

O método REML divide os dados em duas partes: contrastes dos efeitos fixos; e contrastes dos erros (isto é, todos os contrastes com esperança zero) os quais contém informações somente sobre os componentes de variância. Apenas os contrastes dos erros são então usados para estimar os componentes de variância, uma vez que eles contém todas as informações disponíveis sobre os parâmetros de variância. Isto é feito pela projeção dos dados no espaço residual ou espaço vetorial dos contrastes dos erros. Os dados projetados tem Log L dado por $-2RL = [N - r(X)]\log 2\pi - \log\left|X'X\right| + \log\left|XV^{-1}X\right| + \log\left|V\right| + (y - X\hat{b})'V^{-1}(y - X\hat{b})$, em que N é o número de dados e r(X) é o posto da matriz de incidência dos efeitos fixos. Os componentes de variância são então estimados pela maximização do logaritmo da função RL dos dados projetados.

O Log L dos dados originais é dado por $-2L = N\log 2\pi\left| + \log\right|V\left| + (y - Xb)'V^{-1}(y - Xb)\right.$. A função RL tem termos adicionais em relação a L. O único termo adicional relevante para a estimação de componentes de variância é $\log\left|XV^{-1}X\right|$, o qual efetivamente remove os graus de liberdade usados na estimação dos efeitos fixos. Essa diferença entre RL e L reflete exatamente a diferença entre REML e ML.

A idéia inicial de maximizar a parte da função de verossimilhança que é invariante aos parâmetros de locação do modelo, isto é, aos efeitos fixos, foi de Anderson e Bancroft (1952) e Thompson Jr. (1962), para o caso de delineamentos balanceados. Mas, em termos de uma base mais geral abrangendo o caso desbalanceado e de uma descrição matemática formal, é devido a Patterson e Thompson. Também Robertson (1962), considerando a estimação de componentes de variância a partir de dados desbalanceados em modelos de classificação simples (única), derivou pesos ótimos para as médias de família (de acordo com seus tamanhos), os quais são, efetivamente, aqueles usados na estimação REML. O procedimento REML é também denominado de máxima verossimilhança marginal e, na Europa, por máxima verossimilhança residual. Segundo Thompson (2003, comunicação pessoal), o nome correto e original é máxima verossimilhança residual pois, de fato, maximiza-se uma função de verossimilhança dos resíduos, ou seja, empregando não os dados observados y, mas sim os resíduos $(y - X\hat{b})$, em que X é a matriz de incidência dos efeitos fixos.

Os métodos ML e REML requerem solução iterativa, pela não linearidade das equações, fato que dificulta a derivação de estimadores explícitos. Assim, os componentes de variância iteragem nas equações de modelo misto do BLUP até a convergência para um valor adequado. Na maioria das vezes, a estimação dos componentes de variância e a predição de valores genéticos são realizados, simultaneamente, pelo procedimento REML/BLUP. Neste caso, o procedimento BLUP é denominado BLUP empírico.

Os métodos de máxima verossimilhança permitem, sob certas condições, levar em conta os efeitos da seleção. É necessário que todas as informações que tenham contribuído para a seleção sejam incluídas na análise (bem como o parentesco entre os indivíduos), exceto se essas informações não forem correlacionadas com o caráter sob análise. Mesmo se essas condições forem apenas parcialmente obedecidas, o método fornece estimativas menos tendenciosas que aquelas obtidas pelo método III, de Henderson, ou outros métodos baseados em análise de variância (Meyer, 1989a). De maneira geral, argumentos teóricos e evidências indicam que inferências pontuais realizadas a partir de funções de verossimilhança não são afetadas por algumas formas de seleção (Gianola et al., 1989). Essa propriedade fez com que o método REML se tornasse padrão para a estimação de componentes de variância em programas de melhoramento genético.

O método REML elimina o vício devido a mudanças nas freqüências alélicas pela seleção, pelo uso da matriz de parentesco completa (A). Assim, torna-se possível a obtenção de componentes de variância para uma população base não selecionada e a predição de valores genéticos de indivíduos de quaisquer gerações é realizada com precisão. O uso da matriz A completa leva em consideração as alterações na variância genética devida à endogamia e ao desequilíbrio de ligação, as quais resultam da seleção (Kennedy e Sorensen, 1988) e, considera também, a tendência genética ou ganho genético realizado.

Os métodos REML e ML exigem normalidade para que os estimadores tenham propriedades desejáveis. Entretanto, tais estimadores podem ser robustos aos desvios da normalidade, gerando estimativas razoáveis mesmo quando a forma da distribuição não é especificada (Harville, 1977). O método LS (quadrados mínimos ou análise de variância) apresenta a propriedade de não vício independentemente da normalidade ou não dos dados (Lynch e Walsh, 1997).

Para dados balanceados, o estimador REML (negligenciando a suposição de normalidade e a restrição de estimativas REML ao espaço paramétrico) é idêntico ao estimador LS (análise de variância), o qual apresenta as propriedades desejáveis BUE (melhor estimador não viciado) e BQUE (melhor estimador quadrático não viciado) sob normalidade (Searle et al., 1992).

O método REML é uma ferramenta flexível para a estimação de componentes de variância e efeitos fixos, predição de efeitos aleatórios tais quais os valores genéticos e análise estatística em geral. Apresenta as seguintes vantagens:

- Pode ser aplicada a dados desbalanceados.
- É uma generalização da ANOVA para contemplar situações mais complexas e também pode ser derivado sob o enfoque bayesiano, fatos que confirmam a sua generalidade como procedimento ótimo.
- Permite ajustar modelos e delineamentos que não podem ser acomodados pela ANOVA.
- Permite o ajuste de vários modelos alternativos, podendo-se escolher o que se ajusta melhor aos dados e, ao mesmo tempo, é parcimonioso (apresenta menor número de parâmetros).
- Permite lidar com estruturas complexas de dados (medidas repetidas, diferentes anos, locais e delineamentos).
- Permite utilizar simultaneamente um grande número de informações, provenientes de diferentes gerações, locais e idades, gerando estimativas e predições mais precisas.
- No melhoramento animal não exige dados obtidos sob estruturas rígidas de experimentação, podendo ser aplicada a dados obtidos normalmente nos programas de melhoramento, os quais não precisam estar associados a delineamentos, bastando que se tenha informações sobre a genealogia dos indivíduos.
- Permite a estimação dos efeitos de dominância e epistáticos, além dos aditivos, pois utiliza maior número de relações de parentesco.
- Permite comparar indivíduos através do tempo e do espaço.

- Permite a simultânea correção para os efeitos ambientais, estimação de componentes de variância e predição de valores genéticos.
- Permite maior flexibilidade na modelagem, contemplando plenamente a análise de dados correlacionados devido ao parentesco, distribuição temporal e espacial.

O procedimento REML requer que y tenha distribuição normal multivariada. Entretanto, vários autores relatam que os estimadores REML são também apropriados quando não se verifica normalidade dos dados (Harville, 1977; Meyer, 1989).

5 ESTIMAÇÃO BAYESIANA DE COMPONENTES DE VARIÂNCIA E RELAÇÃO COM ML E REML

A inferência estatística bayesiana baseia-se na distribuição condicional do parâmetro (θ) dado o vetor de dados (y), ou seja, na distribuição a posteriori do parâmetro dadas as observações fenotípicas. O princípio bayesiano é atribuído postumamente a Thomas Bayes, que nunca publicou em vida um trabalho matemático. No entanto, Stigler (1986) relata que a base desse princípio foi publicada antes por Saunderson (1683-1739), um cego professor de ótica, que publicou vários artigos matemáticos. O Teorema de Bayes, definido em termos de densidades de probabilidade, tem a seguinte formulação para a distribuição de uma variável aleatória contínua:

$$f(\theta|y) = \frac{f(y|\theta)\ f(\theta)}{\int_R f(y|\theta)\ f(\theta)\ d\theta}.$$

θ: vetor de parâmetros

y: vetor de dados ou de informações obtidas por amostragem

$f(\theta/y)$: distribuição condicional de θ dado y, ou distribuição a posteriori (que é a base da estimação e predição bayesiana).

$f(y/\theta)$: função densidade de probabilidade da distribuição condicional de uma observação (y) dado θ (denominada função de verossimilhança ou modelo para os dados).

f(θ) : função densidade de probabilidade da distribuição a priori, que é também a densidade marginal de θ. Esta função denota o grau de conhecimento acumulado sobre θ, antes da observação de y.

f(y /θ) f(θ): função densidade conjunta de y e θ.

$f(y) = \int_R f(y,\theta)\ d\theta = \int_R f(y|\theta)\ f(\theta)\ d\theta = E_\theta\left[f(y|\theta)\right]$ - distribuição marginal ou preditiva de y com respeito a θ, onde R é a amplitude da distribuição de θ. E_θ significa esperança com respeito à distribuição de θ. (A integração da distribuição conjunta, no espaço paramétrico, produz a marginal de y).

Verifica-se, então, que a distribuição a posteriori é proporcional à verossimilhança x *priori*, ou seja, a função de verossimilhança conecta a priori à posteriori usando para isto os dados do experimento (observações). Dessa forma, a distribuição a posteriori contempla o grau de conhecimento prévio sobre o parâmetro (θ) e também as informações adicionais propiciadas pelo experimento (y).

No contexto dos modelos lineares mistos, os valores genéticos (θ_1) são preditos simultaneamente à estimação dos efeitos fixos (θ_2) e dos componentes de variância (θ_3). Na abordagem bayesiana, a avaliação genética pode ser obtida, de maneira geral, pela construção da densidade a *posteriori* f (θ_1, θ_2, θ_3|y) e, se necessário, pela integração de f (θ_1, θ_2, θ_3|y) em relação a θ_2 e θ_3. Estes (θ_2 e θ_3) são denominados parâmetros de perturbação (*nuisance)* e, por isso, devem ser integrados fora, exceto θ_2 em alguns casos, onde o mesmo constitui-se em uma parte integrante da função de mérito total (neste caso, a função de mérito depende da combinação linear de θ_1 e θ_2).

A obtenção de θ_1 requer a integração ou o conhecimento de θ_2 e θ_3. Henderson (1973) propôs o método BLUP para situações em que θ_3 é conhecido e θ_2 não o é. Para situações em que θ_3 não é conhecido, este autor sugeriu que o procedimento de máxima verossimilhança (ML) propiciaria estimativas razoáveis. Conforme Gianola e Fernando (1986), argumentos bayesianos, que não requerem normalidade e linearidade, permitem validar a intuição de Henderson.

A distribuição de θ_1, θ_2, e θ_3, dado y é proporcional a

$f(\theta_1, \theta_2, \theta_3|y) \alpha f(y|\theta_1, \theta_2, \theta_3) . f(\theta_1, \theta_2, \theta_3)$

Concentrando o interesse em θ_1 (o vetor de valores genéticos), deve-se integrar θ_2 e θ_3 por meio de

$$f(\theta_1 | y) = \int_{R_{\theta_2}} \int_{R_{\theta_3}} f(\theta_1 | \theta_2, \theta_3, y) \cdot f(\theta_2, \theta_3 | y) \cdot d\theta_2 \, d\theta_3 .$$

Tomando a distribuição conjunta a *posteriori* de forma que a maioria da densidade esteja na moda $(\hat{\theta}_2, \hat{\theta}_3)$, tem-se:

$$f(\theta_1 | y) \doteq f(\theta_1 | \theta_2 = \hat{\theta}_2, \theta_3 = \hat{\theta}_3, y) .$$

Usando prioris não informativas para θ_2 e θ_3, tem-se que $\hat{\theta}_2$ e $\hat{\theta}_3$ são precisamente estimadores ML de θ_2 e θ_3, pois neste caso $f(\theta_2, \theta_3|y) \alpha f(y|\theta_2, \theta_3)$, ou seja a densidade de θ_2 e θ_3 dado y é proporcional à função de verossimilhança, de forma que a moda da *posteriori* conjunta corresponde ao máximo da função de verossimilhança, produzindo estimadores ML.

Uma abordagem alternativa para inferência sobre θ_1 consiste em obter $f(\theta_1, \theta_2 | y) \doteq f(\theta_1, \theta_2 | \theta_3 = \hat{\hat{\theta}}_3, y)$, onde $\hat{\hat{\theta}}_3$ refere-se à moda da densidade marginal de θ_3, dado y. Para obtenção de $\hat{\hat{\theta}}_3$ deve-se integrar θ_2 em $f(\theta_2, \theta_3 | y) \alpha f(y | \theta_2, \theta_3)$ e então maximizar $f(\theta_3|y)$. Usando-se uma priori não informativa para θ_3, sob normalidade $\hat{\hat{\theta}}_3$ é um estimador de máxima verossimilhança restrita (REML) para θ_3 (Harville, 1977). Assim, se o interesse reside na inferência conjunta para θ_1 e θ_2 basta usar $f(\theta_1, \theta_2 | y) \doteq f(\theta_1, \theta_2 | \theta_3 = \hat{\hat{\theta}}_3, y)$, que sob normalidade é equivalente à solução BLUP das equações de modelo misto com θ_3 substituído pelas estimativas REML de θ_3 (desde que se tenha usado prioris não informativas para θ_2 e θ_3).

Inferências sobre componentes de variância devem ser baseadas em $f(\theta_3|y) \alpha f(y|\theta_3) . f(\theta_3)$, em que θ_3 contém variâncias e, portanto, $f(\theta_3|y)$ é definida na amplitude $(0, \infty)$ para cada um dos elementos de θ_3, de forma que nunca surgem problemas de estimativas negativas de componentes de variância (Box e Tiao, 1973). $f(\theta_3|y)$ é obtida integrando-se θ_1 em $f(\theta_1, \theta_2, \theta_3|y)$, produzindo $f(\theta_2,$

$\theta_3|y$) e integrando-se θ_2 nesta última. Neste caso, f(θ_2, $\theta_3|y$) conduz aos estimadores ML de θ_2 e θ_3 e f($\theta_3|y$) conduz a um estimador REML de θ_3. Segundo Gianola e Fernando (1986), isto (eliminação das influências de θ_2 ou dos efeitos fixos) mostra precisamente porque REML deve ser preferido em relação a ML, ou seja, estes argumentos são mais fortes do que os apresentados por Patterson e Thompson (1971), que enfatizaram a propriedade de vício do ML.

Em suma, REML e estimativas Bayesianas de componentes de variância quando se usam distribuições a priori não informativas para os componentes de variância, são procedimentos idênticos. Gianola (2001) relata que a derivação Bayesiana de REML é o maior confirmador da eficácia desse método. Revisões atualizadas dos métodos REML e Bayesiano são apresentadas pelos dois maiores especialistas em estimação de componentes de variância na atualidade (Gianola, 2002; Thompson, 2002). Aplicações e comparações práticas dessas duas abordagens no melhoramento vegetal no Brasil são apresentadas por Resende (1997; 1999; 2000b; 2002) e Resende et al. (2001a).

Segundo Thompson et al. (2005), uma razão para o uso da abordagem bayesiana é que ela permite o uso natural de informação a priori dos parâmetros genéticos. Entretanto, tal autor comenta que dentre mais de 30 trabalhos usando a abordagem bayesiana, apresentados no Congresso Mundial de Genética e Melhoramento Animal, em 2002, 83 % não quantificaram conhecimento *a priori* ou usaram conhecimento *a priori* vago para os parâmetros genéticos. Dessa forma, a abordagem usada nada mais é do que a obtenção de REML via métodos de amostragem (cadeias de Markov e Monte Carlo - MCMC). Nesse sentido, Harville (2004) propõe o uso da amostragem de Gibbs como uma forma de tornar o REML computacionalmente possível para grande conjuntos de dados e modelos complexos. Também Thompson et al. (1994) discutem idéias nesse sentido. Maiores detalhes sobre esse tema são apresentados no tópico 13.

6 MÁXIMO E CURVATURA DA VEROSSIMILHANÇA

Além da utilidade como método de estimação paramétrica, por meio de seu ponto de máximo, a verossimilhança é uma ferramenta adequada para lidar com a incerteza contida em conjuntos de dados de tamanho limitado. Nesse caso, é a função de verossimilhança completa que

contempla toda a informação contida nos dados a respeito do parâmetro de interesse e não apenas o maximizador da verossimilhança. O ponto de máximo é apenas um número e então tal ponto dificilmente será suficiente para representar a função.

Quando uma função quadrática representa bem a função Log L em torno de seu máximo, torna-se necessário ao menos duas quantidades para representar a função de verossimilhança. Essas quantidades são a locação de seu máximo e a curvatura no ponto de máximo. Nesse caso, a função de verossimilhança é denominada ***regular*** e, de maneira geral, as funções de verossimilhança se tornam regulares com o aumento do tamanho da amostra. Esta aproximação quadrática é fundamental para o cálculo no contexto da verossimilhança.

Função Escore

A ***função escore*** é definida como a derivada primeira da função Log L e, considerando θ um parâmetro escalar, é dada por:

$S(\theta) = \frac{\partial}{\partial\theta} LogL(\theta)$, em que ∂ refere-se ao operador de derivação ou diferenciação parcial de primeira ordem.

O ***estimador*** $\hat{\theta}$ ***de máxima verossimilhança*** de θ ou o ponto de máximo da função Log L(θ) é a solução da função escore dada por $S(\theta) = 0$. No caso multiparamétrico, o ponto de máximo da função é aquele propiciado pelas estimativas mais plausíveis, considerando todos os parâmetros simultaneamente.

No ponto de máximo $\hat{\theta}$, a derivada segunda de Log L é negativa e a ***curvatura*** no ponto $\hat{\theta}$ é definida como $I(\hat{\theta})$ em que $I(\theta) = -\frac{\partial^2}{\partial\theta^2} LogL(\theta) = -\frac{\partial}{\partial\theta} S(\theta)$, em que ∂^2 refere-se ao operador de derivação parcial de segunda ordem. A última igualdade revela que a curvatura equivale a menos a inclinação da função escore.

Informação Observada de Fisher

Informação é definida como menos a derivada segunda da função de verossimilhança e ***informação observada*** é definida como menos a derivada segunda da função de verossimilhança tomada no ponto de máximo. Uma grande curvatura $I(\hat{\theta})$ está associada com um pico afiado, indicando intuitivamente menos incerteza sobre o parâmetro. Na teoria de verossimilhança, $I(\hat{\theta})$ é uma quantidade essencial denominada ***Informação Observada de Fisher (IOF)***. A IOF é um número e não uma função, pois ela é avaliada no ponto de máximo.

Verossimilhança, Função Escore, Máximo e Informação Observada de Fisher para a Distribuição Normal

Considerando y_1,..., y_n uma amostra aleatória independente e identicamente distribuída obtida de uma distribuição normal $N(\theta,\sigma^2)$, assumindo σ^2 conhecido e ignorando termos constantes irrelevantes, tem-se que o Log da função de verossimilhança é dado por $LogL(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2$

A função escore para a média θ é dada por:

$$S(\theta) = \frac{\partial}{\partial\theta} LogL(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)$$

Igualando essa derivada primeira a zero tem-se $\hat{\theta} = \bar{y}$ como estimador de máxima verossimilhança de θ.

A derivada segunda da função Log L fornece a IOF dada por $I(\hat{\theta}) = \frac{n}{\sigma^2}$.

A variância da estimativa de máxima verossimilhança é dada por $Var(\hat{\theta}) = \frac{\sigma^2}{n} = I^{-1}(\hat{\theta})$ de forma que quanto maior a informação de Fisher, menor é variância da estimativa do parâmetro. O erro padrão da estimativa $\hat{\theta}$ é dado por $s(\hat{\theta}) = \frac{\sigma}{\sqrt{n}} = I^{-1/2}(\hat{\theta})$.

Verossimilhança, Função Escore, Máximo, Curvatura e Informação Observada de Fisher para a Distribuição Binomial

Considerando y com distribuição binomial $Bin(n,\theta)$, o Log da função de verossimilhança é dado por $LogL(\theta) = y \log \theta + (n-y) \log(1-\theta)$

A função escore para a média θ é dada por

$$S(\theta) = \frac{\partial}{\partial\theta} LogL(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

Igualando essa derivada primeira a zero tem-se $\hat{\theta} = y/n$ como estimador de máxima verossimilhança de θ.

A derivada segunda da função Log L fornece a curvatura dada por $I(\theta) = -\frac{\partial^2}{\partial\theta^2} LogL(\theta) = \frac{y}{\theta^2} + \frac{n-y}{(1-\theta^2)}$.

No ponto de máximo, a IOF é dada por $I(\hat{\theta}) = \frac{n}{\hat{\theta}(1-\hat{\theta})}$.

A variância da estimativa de máxima verossimilhança é dada por $Var(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n} = I^{-1}(\hat{\theta})$.

Estatística Escore e Informação Esperada de Fisher

$S(\theta)$ como função de θ é denominada função escore e como função de uma variável aleatória para θ fixado é denominada ***estatística escore***. A estatística escore tem ricas propriedades freqüentistas. A distribuição de amostragem da função escore mostra o que se deve esperar quando os dados variam de amostra para amostra.

A quantidade IOF varia de amostra para amostra e a ***Informação Esperada de Fisher*** (IEF) é definida como $\Gamma(\theta) = E_\theta I(\theta)$. O valor esperado é tomado no ponto fixo e verdadeiro de θ. Verdadeiro no sentido de que os dados são gerados sob aquele valor de θ.

Existem diferenças qualitativas entre IOF e IEF. A IEF tem significado como uma função de θ através dos valores admissíveis de θ, mas a IOF tem significado somente na vizinhança de $\hat{\theta}$. Ou seja, a IOF aplica-se apenas a um conjunto de dados já que é uma quantidade observada da verossimilhança, devendo ser interpretada como uma estatística única e não como uma função. Por outro lado, a IEF é uma quantidade média sobre todos os possíveis conjuntos de dados gerados sob o valor verdadeiro do parâmetro. Assim, não é óbvio se a IEF é uma medida de informação adequada para um determinado conjunto de dados. A IEF informa sobre a dificuldade de se estimar θ, ou seja, parâmetros com maior IEF podem ser estimados mais facilmente e requerem menos amostras para que sejam estimados com precisão. A IEF equivale à variância da estatística escore.

7 DERIVAÇÃO DO MÉTODO DE MÁXIMA VEROSSIMILHANÇA SOB MODELOS MISTOS

O objetivo dos métodos ML e REML é encontrar um conjunto de parâmetros que maximizam a verossimilhança dos dados. A verossimilhança dos dados para um determinado modelo pode ser escrito como uma função. Segundo os fundamentos de cálculo matemático, para encontrar o máximo dessa função, deve-se tomar a primeira derivada ou diferencial dessa função e igualar o resultado a zero. Isto propicia o conjunto de parâmetros que conduzem a função a um ponto crítico máximo, desde que não se tenha atingido um ponto de mínimo. Isto pode ser verificado usando o sinal da derivada segunda. Sinal positivo da derivada segunda indica concavidade para cima, ou seja, ponto de mínimo. Sinal negativo da derivada segunda indica concavidade para baixo, ou seja, ponto de máximo. A seguir é apresentado, passo a passo, a obtenção das funções de verossimilhança e de verossimilhança residual a serem maximizadas no contexto dos modelos mistos ou modelos de componentes de variância. Os procedimentos de maximização são apresentados no tópico 13.

Função Densidade de Probabilidade da Variável Aleatória y

Assumindo uma variável aleatória y com distribuição normal com média μ e variância σ^2, ou seja, $y \sim N(\mu, \sigma^2)$, tem-se a função densidade de probabilidade dada por

$$f(y|\mu,\sigma^2) = \frac{1}{[2\pi\sigma^2]^{1/2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right\}$$

Considere o modelo misto geral da forma $y = Xb + Za + e$ com y tendo distribuição normal multivariada com as seguintes distribuições e estruturas de médias e variâncias (Henderson, 1984, Searle et al., 1992; Brown e Prescott, 2001; Thompson, 2002; Thompson et al., 2003; Thompson, 2005, Thompson et al., 2005):

$a \sim N(0,\ G)$ $\qquad$ $E(y) = Xb$

$e \sim N(0,\ R)$ $\qquad$ $Var(y) = V = ZGZ' + R$

em que:

y: vetor de observações.

b: vetor paramétrico dos efeitos fixos, com matriz de incidência *X*.

a: vetor paramétrico dos efeitos aleatórios, com matriz de incidência Z.

e: vetor de erros aleatórios.

G: matriz de variância – covariância dos efeitos aleatórios.

R: matriz de variância – covariância dos erros aleatórios.

0: vetor nulo.

Neste caso, a função densidade de probabilidade de y é dada por

$$f(y|Xb,V) = \frac{1}{2\pi^{(1/2)N}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y-Xb)'V^{-1}(y-Xb)\right\}$$

Essa função fornece a probabilidade de se observar y dado os parâmetros b e V.

Função de Verossimilhança dos Parâmetros Dado y

O interesse agora é na probabilidade de que y tenha sido observado com determinados valores dos parâmetros b e V, ou seja, dado y qual é a verossimilhança dos parâmetros b e V. Logo, com y fixado, o interesse é encontrar quais parâmetros tornam a função $f(b,V|X,y)$ com valor máximo. Essa função é a função de verossimilhança e pode ser escrita como

$$L(b,V|X,y) = \frac{1}{2\pi^{(1/2)N}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y-Xb)'V^{-1}(y-Xb)\right\}.$$

Logaritmo da Função de Verossimilhança dos Parâmetros Dado y

Tomando-se o logaritmo natural na expressão acima obtém-se a função Log L, de verossimilhança dos parâmetros b e V dados a matriz de incidência X e os dados observados (y):

$$Log\ L(b,V|X,y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(y-Xb)'V^{-1}(y-Xb)$$

Obtenção das Derivadas Parciais de Primeira Ordem da Função Log L

O próximo passo é a derivação dos estimadores ML a partir da maximização da função Log L. Nesse ponto, intenso uso das regras de ***Derivação Matricial*** é feito. Essas regras são detalhadamente descritas por Searle (1982) e Harville (1997).

A derivação com respeito ao vetor de efeitos fixos é dada por

$$\frac{\partial Log\ L(b,V|X,y)}{\partial b} = -\frac{1}{2}\frac{\partial (y-Xb)'V^{-1}(y-Xb)}{\partial b} = X'V^{-1}(y-Xb)$$

A derivação com respeito aos componentes de variância (denominados σ_k^2 para referir-se ao fator aleatório k) é dada por

$$\frac{\partial LogL(b,V|X,y)}{\partial \sigma_k^2} = -\frac{1}{2}tr(V^{-1}V_k) + \frac{1}{2}(y-X\hat{b})'V^{-1}V_kV^{-1}(y-X\hat{b}) + \frac{1}{2}(b-\hat{b})X'V^{-1}V_kV^{-1}X(b-\hat{b})$$

Obtenção dos Estimadores ML dos Parâmetros em b e V

Igualando as equações acima a zero e resolvendo, têm-se os estimadores ML.

a) Estimador ML para os efeitos fixos:

$$\frac{\partial Log\ L(b,V|X,y)}{\partial b} = X'V^{-1}(y-Xb) = 0$$

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

Esse é o estimador ML e também BLUE ou GLS de b. Quando a inversa de $(X'V^{-1}X)$ não existe, pode-se empregar uma inversa generalizada.

b) Estimador ML para os componentes de variância:

$\frac{\partial Log L(b,V|X,y)}{\partial \sigma_k^2} = -\frac{1}{2}tr(V^{-1}V_k) + \frac{1}{2}(y-X\hat{b})'V^{-1}V_kV^{-1}(y-X\hat{b}) + \frac{1}{2}(b-\hat{b})X'V^{-1}V_kV^{-1}X(b-\hat{b}) = 0$ Substituindo b por $\hat{b}$ na expressão acima, o último termo torna-se zero. Rearranjando a igualdade, tem-se:

$$tr(V^{-1}V_k) = (y-X\hat{b})'V^{-1}V_kV^{-1}(y-X\hat{b})$$

Com base na definição do projetor ortogonal $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$, a expressão acima pode ser escrita como: $tr(V^{-1}V_k) = y'\hat{P}V_K\hat{P}y$, em que P é denotada como $\hat{P}$ para significar que a mesma é função de V e depende dos componentes de variância que ainda precisam ser estimados.

Para aplicações em genética sob modelo individual univariado, G = Aσ_a^2 e R =Iσ_e^2 e as equações para os componentes de variância são: $tr(V^{-1}I\sigma_e^2) = y'\hat{P}I\sigma_e^2\hat{P}y$, ou seja, $tr(V^{-1}) = y'\hat{P}\hat{P}y$ para o componente residual; $tr(V^{-1}ZAZ'\sigma_a^2) = y'\hat{P}ZAZ'\sigma_a^2\hat{P}y$, ou seja $tr(V^{-1}ZAZ') = y'\hat{P}ZAZ'\hat{P}y$ para o componente genético.

Estas equações para $\hat{b}$ e para os componentes de variância não apresentam solução direta ou explícita pois: (i) $\hat{b}$ depende de $\hat{V}$, que ainda precisa ser estimada; (ii) os componentes de variância que se deseja estimar em $\hat{V}$ encontram-se em ambos os lados das equações para $\hat{\sigma}_a^2$ e $\hat{\sigma}_e^2$. Assim, o conjunto de todas as equações representa funções não lineares, e métodos iterativos devem ser usados para obtenção de estimativas ML para os três parâmetros em questão. Esses métodos iterativos são apresentados no tópico 13.

8 DERIVAÇÃO DO MÉTODO DE MÁXIMA VEROSSIMILHANÇA RESIDUAL SOB MODELOS MISTOS

O método REML divide os dados e a função densidade de probabilidade em duas partes: contrastes dos efeitos fixos; e contrastes dos erros ou resíduos (isto é, todos os contrastes com esperança zero) os quais contém informações somente sobre os componentes de variância. Assim, a função de verossimilhança residual usa apenas os contrastes dos erros, denotados por K'y, em que K'X = 0 e K' tem posto N – r(X). Dessa forma, o REML é baseado em uma transformação linear de y, tal que os efeitos fixos são removidos do modelo. A matriz K estabelece os contrastes envolvendo os componentes aleatórios das observações.

O modelo misto agora tem a forma y* = $K'y = K'Xb + K'Za + Ke = K'Za + Ke$. Nota-se que o modelo agora é só de componentes de variância, sem efeitos fixos. E a matriz V agora é dada por V* = KVK'.

Função de Verossimilhança Residual dos Parâmetros Dado y

A função de verossimilhança residual pode ser escrita como

$$L(V|K', y) = \frac{1}{2\pi^{(1/2)[N-r(X)]}|K'VK|^{1/2}} \exp\left\{-\frac{1}{2}(K'y)'(K'VK)^{-1}(K'y)\right\}.$$

Por ser uma função dos contrastes residuais, essa função nada informa a respeito dos efeitos fixos b.

Logaritmo da Função de Verossimilhança Residual dos Parâmetros Dado y

Tomando-se o logaritmo natural na expressão acima, obtém-se a função Log L, de verossimilhança residual dos parâmetros em V dado K'y:

$$Log\, L(V|K', y) = -\frac{N - r(X)}{2}\log(2\pi) - \frac{1}{2}\log|K'VK| - \frac{1}{2}y'K(K'VK)^{-1}(K'y).$$

O primeiro termo dessa expressão é constante e não depende dos componentes de variância desconhecidos. Assim, podem ser ignorados. Searle et al. (1992) demonstram que $\log|K'VK| = \log|V| + \log|X'V^{-1}X|$ e que $y'K(K'VK)^{-1}(K'y) = y'Py = (y - X\hat{b})'V^{-1}(y - X\hat{b})$.

Assim, a função Log L pode ser escrita como

$$LogL = -0.5\log|V| - 0.5\log|X'V^{-1}X| - 0.5(y - X\hat{b})'V^{-1}(y - X\hat{b}).$$

O termo $\left|X'V^{-1}X\right|$ está presente no REML mas não aparece na derivação do ML e é referido como função de penalização porque os efeitos fixos não são conhecidos.

Usando resultados da álgebra matricial, demonstra-se que:

$\log\left|V\right| = \log\left|R\right| + \log\left|G\right| + \log\left|G^{-1} + Z'R^{-1}Z\right|$ e que

$\log\left|X'V^{-1}X\right| = \log\left|C\right| - \log\left|G^{-1} + Z'R^{-1}Z\right|$. Combinando essas igualdades, tem-se

$LogL = -0.5\log\left|R\right| - 0.5\log\left|G\right| - 0.5\log\left|C\right| - 0.5y'Py$.

A mesma função Log L pode também ser expressa como:

$$LogL = -0.5[N - r(X) - q]\log\sigma_e^2 - 0.5q\log\sigma_a^2 - 0.5\log\left|C\right| - 0.5y'Py$$

A matriz C acima é dada por:

$$C = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\sigma_e^2 \end{bmatrix}^{-1}$$

Uma dessas formas alternativas de Log L é que deve ser maximizada visando a obtenção do REML.

Outra forma da função de verossimilhança restrita a ser maximizada é dada por (Patterson e Thompson, 1971; Searle et al., 1992, Thompson, 1973; 1977; 1980; 2002; Thompson e Welham, 2003):

$$L = -\frac{1}{2}\left(\log\left|XV^{-1}X\right| + \log\left|V\right| + v\ \log\ \sigma_e^2 + y'Py/\sigma_e^2\right)$$

$$= -\frac{1}{2}\left(\log\left|C^*\right| + \log\left|R\right| + \log\left|G\right| + v\ \log\ \sigma_e^2 + y'Py/\sigma_e^2\right)$$

em que:

$V = R + ZGZ'$; $\quad P = V^{-1} - V^{-1}X\ (X'V^{-1}X)^{-1}X'V^{-1}$;

v = N-r(x): graus de liberdade, em que *N* é o número total de dados e *r(x)* é o posto da matriz *X*;

*C**: matriz dos coeficientes das equações de modelo misto.

Para modelos univariados que envolvem dois fatores aleatórios (além do resíduo) e três componentes de variância a serem estimados, a função *L* a ser avaliada equivale a:

$L = -\frac{1}{2}\left[(N - r(x) - N_a - N_c)\, log_e\, \hat{\sigma}_e^2 + log_e|C| + N_a\, log_e\, \hat{\sigma}_a^2 + log_e|A| + N_c\, log_e\, \sigma_c^2 + y'Py / \sigma_e^2\right]$, em que:

$|C|$: determinante de uma submatriz não singular de *C* com posto máximo, em que *C* é a matriz dos coeficientes das equações de modelo misto.

N_a: número de níveis no efeito aleatório *a*.

N_c: número de níveis no efeito aleatório c.

Neste caso, *V* (em *P*) refere-se a todos os efeitos aleatórios. Também $log_e|A|$ não depende dos parâmetros a serem estimados, ou seja, é constante e não necessita ser calculada para maximizar *L*.

Obtenção das Derivadas Parciais de Primeira Ordem da Função Log L

O próximo passo é a derivação dos estimadores REML a partir da maximização da função Log L.

Sendo b* = 0 no modelo y*, a derivação com respeito aos componentes de variância (denominados σ_k^2 para referir-se ao fator aleatório k) é dada por

$$\frac{\partial Log\ L(V|K', y)}{\partial \sigma_k^2} = -\frac{1}{2}\, tr((V^*)^{-1} V_k^*) + \frac{1}{2}(y^*)'(V^*)^{-1} V_k^* (V^*)^{-1} y^*$$

Obtenção dos Estimadores REML dos Parâmetros em V

Igualando as equações acima a zero e resolvendo, tem-se os estimadores REML para os componentes de variância.

As equações para os componentes de variância são:

$tr(\hat{P}) = y'\hat{P}\hat{P}y$ para o componente residual.

$tr(\hat{P}ZAZ') = y'\hat{P}ZAZ'\hat{P}y$ para o componente genético.

Essas equações têm a mesma forma das equações para ML, porém V^{-1} é substituída por P. Isto decorre do fato de que, na derivação do ML, os efeitos fixos são assumidos como paramétricos ou conhecidos sem erro. Por outro lado, no REML, a fração $\left|X'V^{-1}X\right|$ na função Log L da parte aleatória das observações penaliza pelo fato de não se conhecer b e ao mesmo tempo considera a perda de graus de liberdade em sua estimação. Uma diferença fundamental é que o REML para os componentes de variância não produz estimativas de b, visto que o método inicialmente remove os efeitos fixos (fazendo b* = 0,) antes de estimar os componentes de variância.

Assim, estimativas de b podem ser obtidas pela maximização da parte da função de verossimilhança referentes aos efeitos fixos. O Log L dessa função é dado por

$$Log\ L(b|V,y) = -\frac{r(X)}{2}\log(2\pi) - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{1}{2}\{y'V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}y) - y'V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}Xb + b'X'V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}Xb\}$$

. A maximização dessa função conduz ao estimador GLS ou BLUE de b dado por $\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$.

Da mesma forma, o conjunto de todas as equações representa funções não lineares, e métodos iterativos devem ser usados para obtenção de estimativas REML para os parâmetros em questão. Esses métodos iterativos são apresentados no tópico 13.

A função (*L*) de verossimilhança restrita, expressa em termos do logaritmo, pode ser maximizada (visando obter as estimativas REML dos componentes de variância), empregando-se diferentes algoritmos agrupados de acordo com a ordem das derivadas da função de verossimilhança utilizadas, nas seguintes classes: (i) não derivativo (DF-REML de Graser et al., 1987; Meyer, 1989), baseado em procura direta; (ii) baseado em derivadas parciais de primeira ordem (EM-REML, de Dempster et al. 1977; Henderson, 1984; 1986); (iii) baseado em derivadas parciais de primeira e segunda ordens (AI-REML, de Johnson e Thompson, 1995 e Gilmour et al.,1995). Estes algoritmos geraram as denominações EMREML, DFREML e AIREML. Aplicações iniciais do REML usaram também os métodos Newton-Raphson e dos Escores de Fisher (Patterson e Thompson, 1971; Thompson, 1977). Recentemente, um método EM acelerado (PX-EM) foi apresentado por Foulley & Van Dyk (2000).

9 VEROSSIMILHANÇA PERFILADA MODIFICADA E REML

A definição de verossimilhança contempla modelos multi-paramétricos. Entretanto, muitas vezes o interesse reside em apenas um subconjunto de parâmetros, sendo os demais denominados parâmetros de perturbação (*nuisance*) e participam do modelo apenas para ajudar a descrever melhor a variabilidade. Um caso típico é quando o interesse reside nos componentes de variância e os efeitos fixos são considerados *nuisance*. Nesse caso, é necessário um método para concentrar a verossimilhança em um só parâmetro ou grupo de parâmetros por meio da eliminação do parâmetro de *nuisance*.

A abordagem de verossimilhança para eliminar parâmetros de *nuisance* refere-se a substituir tais parâmetros por suas estimativas de máxima verossimilhança para cada valor fixo do parâmetro de interesse. A verossimilhança resultante é então denominada ***verossimilhança perfilada ou concentrada***.

A abordagem bayesiana elimina todos os parâmetros não interessantes, integrando-os fora da distribuição. Entretanto, a função de verossimilhança não é uma função densidade de probabilidade (ou seja, não integra 1) e não obedece leis de probabilidade. Assim, integrar um

parâmetro em uma função de verossimilhança não tem sentido. No entanto, existe uma analogia entre integração na abordagem bayesiana e o conceito de perfil de verossimilhança modificado relatado na seqüência.

Existe um método genérico de transformação de dados y para (v, w) de forma que a distribuição marginal de v e a distribuição condicional de v dado w depende apenas do parâmetro de interesse. Isso caracteriza o que é denominado ***verossimilhança marginal*** e ***verossimilhança condicional***, respectivamente. No entanto, verossimilhanças marginais e condicionais exatas nem sempre estão disponíveis ou são difíceis de derivar. Uma aproximação para essas pode ser obtida modificando-se o perfil de verossimilhança tradicional para se obter o perfil de verossimilhança modificado (Barndorff-Nielsen, 1983).

O Log L associado ao ***perfil de verossimilhança modificado*** (PVM) é dado por (Pawitan, 2001):

$$PVMLogL(\theta) = \log L_p(\theta) - 0.5\log\left|I(\hat{\eta}_\theta)\right| + \log\left|\partial\hat{\eta}/\partial\hat{\eta}_\theta)\right| \quad ,$$

em que:

$\log L_p(\theta)$: Log L do perfil de verossimilhança tradicional;

$\log\left|I(\hat{\eta}_\theta)\right|$: Log L da informação de Fisher associada às estimativas ML de η para valores fixados de θ; pode ser interpretado como um termo de penalização o qual subtrai do perfil de verossimilhança as informações indesejadas do parâmetro de *nuisance* η.

$\left|\partial\hat{\eta}/\partial\hat{\eta}_\theta)\right|$: termo jacobiano que funciona como uma quantidade de preservação da invariância do PVM com respeito a transformações nos parâmetros de *nuisance*. Essa quantidade é igual a zero quando os parâmetros são ***informação ortogonal.***

No contexto dos modelos mistos, tem-se:

Log L da verossimilhança ordinária:

$$LogL\ (b,V|y) = -0.5\log|V| - 0.5\ (y - Xb)'V^{-1}(y - Xb);$$

Log L do perfil de verossimilhança:

$$LogL_p\ (V|y,\hat{b}) = -0.5\log|V| - 0.5\ (y - X\hat{b})'V^{-1}(y - X\hat{b});$$

Log L do perfil de verossimilhança modificado :

$LogL_{PVM}\ (V|y,\hat{b}) = -0.5\log|V| - 0.5\ (y - X\hat{b})'V^{-1}(y - X\hat{b}) - 0.5\log|X'V^{-1}X|$, considerando b e V como parâmetros informação ortogonal, ou seja, $|\partial\hat{b} / \partial\hat{b}_V)\ |$ = 0.

Verifica-se então que essa última expressão é equivalente àquela derivada para o REML e que $-0.5\log|I(\hat{\eta}_\theta)\ | = -0.5\log|X'V^{-1}X|$. Portanto, o REML se ajusta perfeitamente ao conceito de perfil de verossimilhança modificado. Essa é mais uma forma de derivação do REML, fato que revela, mais uma vez, a sua generalidade como procedimento ótimo.

10 ESCOLHA E COMPARAÇÃO ENTRE MODELOS DE ANÁLISE

10.1 Teste da Razão de Verossimilhança (LRT)

10.1.1 Inferência Probabilística ou Critério de Deviance (DIC)

Além de sua utilidade na estimação, o princípio da verossimilhança também permite comparar a adequabilidade de vários modelos, desde que tenham uma estrutura hierárquica ou aninhada. Considerando-se dois modelos *U* e *V* com máximos das funções de verossimilhança (restrita ou não*) L(U) e L(V)* e correspondentes números de parâmetros n_u *e* n_v, tem-se que menos duas vezes o logaritmo da razão de verossimilhança, *D = -2 Log$_e$ ((L(U))/L(V)),* possui, aproximadamente, distribuição χ^2 com n_v - n_u graus de liberdade (assumindo U como hierárquico ou um caso especial de *V*).

As funções *L* e Log_e *L* apresentam o mesmo máximo. Estas funções podem ser expressas também em termos da função *–2* Log_e *L*, que deve ser minimizada. Para maximização de *L* ou Log_e *L*, valores iniciais dos parâmetros são fornecidos e a cada iteração, novos valores são calculados e substituídos na função *–2* Log_e *L* até que esta seja minimizada (e, de maneira equivalente, Log_e L será maximizada) e os parâmetros convirjam para os valores constantes.

A significância da diferença no ajuste de diferentes modelos aos dados pode ser testada usando o Teste da Razão de Verossimilhança de Wilks (LRT), definido por:

$$\lambda = 2Log_e L(V) - 2Log_e L(U)$$

$$\lambda = 2[Log_e L_{p+1} - Log_e L_p].$$

Assim, basta comparar λ [2 vezes a diferença (modelo com maior número de parâmetros – modelo com menor número p de parâmetros) de Log_e L associados a dois modelos ajustados] com o valor da função densidade de probabilidade (Tabela de χ^2) para determinado número de graus de liberdade e probabilidade de erro (Dobson, 1990). O número de graus de liberdade é definido pela diferença no número de parâmetros ou componentes de variância entre modelos. A distribuição assintótica de λ foi derivada por Wilks (1938).

Quando V é o modelo saturado e equivale a uma exata reprodução dos dados, D é denominado deviance do modelo U (dada por -2 log_e L(U)). Quanto menor a deviance de um modelo, menor os resíduos do mesmo e melhor o modelo. Então, alternativamente, a diferença entre as deviances dos dois modelos ajustados pode ser usada para o teste LRT. Esse teste pode ser usado para comparar modelos desde que os mesmos tenham uma estrutura hierárquica e os mesmos efeitos fixos. Isto permite a comparação de modelos com diferentes fatores aleatórios porém com uma mesma estrutura de efeitos fixos. Para comparação de modelos espaciais, essa estatística pode ser usada para avaliar a ordem do modelo a ser ajustado. Então, é possível testar se um modelo de médias móveis de ordem 2 (MA(2)) tem melhor ajuste que um modelo MA(1) de ordem 1, ou se um modelo ARMA (1,1) é melhor do que um modelo AR(1). Entretanto, o uso do LRT é limitado a modelos ajustados sob o mesmo regime de diferenciação. O teste de modelos com diferentes estruturas de efeitos fixos foi considerado por Welham e Thompson (1997).

10.1.2 Inferência Verossimilhança Pura

Em problemas onde a inferência probabilística exata não está disponível, a função de verossimilhança observada pode ser usada diretamente para inferência. Isto pode ser feito por meio da razão de riscos (*odds ratio*), a qual é a própria razão direta entre os valores da função maximizada por dois conjuntos distintos de valores paramétricos a serem avaliados, ou seja, OD = *(L(U))/L(V)).*

A inferência verossimilhança pura pode ser usada quando a teoria de grandes amostras não for adequada ao caso analisado. Esse é o caso de amostras pequenas com distribuição não normal ou o caso de distribuições complicadas, onde a verossimilhança não é regular (ver a definição de regularidade da verossimilhança no tópico 6).

Uma derivação do OD, muito usada no contexto da genética é o teste do LOD escore. LOD significa "*log of odds ratio*", ou seja, logaritmo na base 10 da razão de riscos (odds ratio). Riscos, no caso, quantificados pela verossimilhança de dois modelos a serem comparados. O LOD é dado por $LOD = Log_{10}\ OD = Log_{10}\ (L(U))/L(V)) = \lambda\ /\ [2\ Log\ (10)] = \lambda\ /\ 4.61$. Portanto, existe uma relação direta entre o LOD e o LRT ou λ, ou seja, $LOD = LRT\ /\ 4.61$. Alternativamente, $LRT = 4.61\ LOD$.

Com base nessa última expressão, pode-se associar valores de LOD e p–valores aproximados do LRT. Os valores críticos (λ) de qui-quadrado nos níveis de significância 10 %, 5 %, 1 % e 0.5 % são 2.71, 3.84, 6.64 e 7.88, respectivamente. Esses valores estão associados aos seguintes LOD´s, dados por $LOD = LRT\ /\ 4.61$: 0.588, 0.833, 1.440 e 1.709, respectivamente. Assim, uma inferência aproximada é de que LOD´s maiores que 1.71 já estão associados a elevados (menores do que 0.5 %) níveis de significância. Um LOD escore de 3 significa que uma hipótese é mil vezes mais plausível que a outra. Neste caso, a inferência é baseada apenas na razão de verossimilhança, sem invocar as propriedades distribucionais dos estimadores de máxima verossimilhança.

Com base nos valores de LOD associados aos níveis de significância de 10 %, 5 %, 1 % e 0.5 %, podem ser obtidos os valores de OD por $OD = 10^{LOD}$. Assim, têm-se os valores de OD: 3.9, 6.8, 27.5 e 51.2, respectivamente. Logo, valores da razão direta entre os valores da função

maximizada por dois conjuntos distintos de valores paramétricos a serem avaliados, das ordens de 4, 7, 28 e 51, estariam aproximadamente associados aos níveis de significância 10 %, 5 %, 1 % e 0.5 %, respectivamente.

10.2 Critério de Informação de Akaike (AIC)

Quando dois modelos aninhados são ajustados, aquele com mais parâmetros apresenta maior log L. Entretanto, esse não é necessariamente o melhor modelo. Isto significa que não se pode comparar diretamente os Log L quando o número de parâmetros varia entre modelos. Além do LRT, outro critério para a seleção de modelos é o Critério de Informação de Akaike (ACI), o qual penaliza a verossimilhança pelo número de parâmetros independentes ajustados. Por esse critério, qualquer parâmetro extra deve aumentar a verossimilhança por ao menos uma unidade para que o mesmo entre no modelo. O AIC é dado por AIC = -2 log L + 2 p, em que p é o número de parâmetros estimados. Menores valores de AIC refletem um melhor ajuste global (Akaike, 1974). Assim, os valores de AIC são calculados para cada modelo e aquele com menor valor de AIC é escolhido como melhor modelo.

O primeiro termo do AIC pode ser interpretado como uma medida de ajuste do modelo e o segundo termo como uma penalização. Desse modo, no caso em que se compara modelos com o mesmo número de parâmetros, necessita-se comparar apenas o Log L. A vantagem do AIC é que as comparações não se limitam a modelos com estrutura hierárquica de fatores, fato que faz do AIC uma ferramenta genérica para a seleção de modelos. Pode ser usado, por exemplo, para a comparação entre modelos com erros, apresentando diferentes distribuições. No entanto, os modelos devem ter a mesma estrutura de efeitos fixos.

Uma comparação entre o AIC e o LRT revela que o AIC é potencialmente mais fraco do que a inferência baseada em probabilidade propiciada pelo LRT. Isto porque não existe como impor um nível de significância para uma diferença observada nos AICs. Assim, o AIC é especialmente útil quando uma inferência baseada em probabilidade não é disponível, por exemplo, quando os modelos não possuem estrutura hierárquica.

Comparado ao LRT, o AIC permitirá ao modelo crescer (p > 1), mas não permitirá muitos parâmetros (> 6), conforme a Tabela a seguir, em que 2p ≥ ao qui-quadrado crítico quando p > 6. Usar o AIC é equivalente a mudar o nível de significância, dependendo do número de parâmetros (Pawitan, 2001). A quantidade 2p é o segundo termo da expressão do AIC e refere-se ao acréscimo no AIC em relação ao LRT.

p	1	2	3	4	5	7	8	10
2p	2	4	6	8	10	14	16	20
Qui-quadrado-5%	3.84	5.99	7.81	9.49	11.07	14.07	15.51	18.31

10.3 Critério de Informação Bayesiano (BIC)

Outra abordagem é o Critério de Informação Bayesiano (BIC) de Schwarz (1978), o qual é dado por BIC = -2 log L + p log v, em que v = N – r(x) é o número de graus de liberdade do resíduo. O BIC é calculado para cada modelo e aquele com menor valor é escolhido como melhor modelo. Pode ser usado quando os modelos não possuem estrutura hierárquica. No entanto, os modelos devem ter a mesma estrutura de efeitos fixos. Logicamente, tanto o LRT, o AIC e o BIC dependem da mesma quantidade básica -2 log L.

10.4 Comparação de Modelos com Diferentes Efeitos Fixos

De maneira diferente do contexto da regressão linear, a diferença entre as deviances de dois modelos hierárquicos de efeitos fixos não propicia um teste estatístico adequado. Isto deve-se ao fato de que a verossimilhança residual é que é maximizada e não a verossimilhança dos dados originais. A verossimilhança residual refere-se à verossimilhança dos dados após projeção no espaço residual e, portanto, dois diferentes modelos quanto aos efeitos fixos referem-se a duas diferentes projeções e, conseqüentemente, correspondem a diferentes conjuntos de dados nos quais

os mesmos fatores aleatórios são estimados. Welham e Thompson (1997) propuseram um teste LRT para avaliar fatores fixos de um modelo contra um sub-modelo hierárquico. O método computa a verossimilhança para o modelo fixo completo da forma usual e a mesma projeção é então usada para o submodelo e os efeitos fixos a serem retirados do submodelo são forçados a zero. Isto propicia Log L calculados para o mesmo conjunto de dados projetados, usando a mesma parte aleatória do modelo mas com alguns efeitos fixos restritos a zero no submodelo. A diferença entre os Log L propicia o LRT da maneira usual e os graus de liberdade da estatística de teste (qui-quadrado) equivale ao número de graus de liberdade dos termos da parte fixa do modelo restritos a zero no submodelo.

11 TESTE DE EFEITOS FIXOS E ALEATÓRIOS NO CONTEXTO DOS MODELOS MISTOS

11.1 Teste de Efeitos Fixos

A análise de dados de experimentos balanceados é tradicionalmente realizada por uma análise de variância na qual a soma de quadrados total dos desvios das observações em relação às suas médias é particionada em soma de quadrados devidas aos efeitos de tratamentos (efeitos fixos) e a outros efeitos aleatórios, tais como o erro experimental. Sob a hipótese de nulidade (sob H_o) dos efeitos de tratamentos, a soma de quadrados de tratamentos segue uma distribuição múltipla (produto dos graus de liberdade por uma combinação linear dos componentes de variância dos efeitos aleatórios e somatório dos quadrados dos efeitos fixos) de uma χ^2. A soma de quadrados associada aos efeitos aleatórios (erro, por exemplo), sob as suposições distribucionais padrões, também segue uma distribuição múltipla de uma χ^2 (Elston, 1998).

Uma vez que a razão entre distribuições χ^2 segue uma distribuição F de Snedecor, hipóteses sobre os efeitos de tratamentos podem ser testadas pela razão entre os quadrados médios dos efeitos de tratamentos e do erro (por exemplo), por meio do teste F. Segundo Elston (1998), alguns dos benefícios de um delineamento experimental, ou seja, do balanceamento, referem-se ao fato da variância de todos os contrastes para um dado fator de tratamento ter a mesma esperança sob a

hipótese da nulidade e tal variância ser tão pequena quanto possível. Por outro lado, a análise de dados que não apresenta essas propriedades de balanceamento é muito menos fidedigna, e os principais problemas são os fatos da decomposição da soma de quadrados não ser única e da não existência de um quadrado médio do erro com igual esperança sob a hipótese da nulidade.

Segundo Henderson (1984), os testes de hipóteses referentes aos efeitos fixos são exatos apenas quando os componentes de variância (variância dos efeitos aleatórios) são conhecidos. Isto porque quando se usam as estimativas destes componentes de variância em lugar de seus valores paramétricos, as somas dos quadrados mencionadas anteriormente não têm distribuição χ^2 ou qualquer outra distribuição tratável. Apenas no caso de um modelo de efeitos fixos (exceto o erro), em que a variância do erro –Var (e) – é definida como $Var(e) = I\sigma_e^2$, as distribuições χ^2 e F mencionadas podem ser asseguradas, quando substitui-se σ_e^2 por sua estimativa $\hat{\sigma}_e^2$ (obtida por máxima verossimilhança restrita – REML). Também, para classificações duplas, com um efeito fixo e outro aleatório, testes exatos para os efeitos fixos podem ser encontrados (Khuri e Littell, 1987; Gallo e Khuri, 1990).

Segundo Henderson (1984), duas possibilidades existem para contornar o problema: (i) estimar os componentes de variância por máxima verossimilhança restrita e invocar o teste da razão de verossimilhança – LRT –, através do qual, sob suposição de normalidade e com amostras grandes, -2 log da razão de verossimilhança aproxima uma distribuição χ^2; (ii) considerar os componentes de variância estimados, como sendo paramétricos (isto é tanto mais realístico quanto maior for o tamanho da amostra) e realizar um teste aproximado usando as distribuições χ^2 (para os quadrados médios) e F (para a razão entre os quadrados médios e conseqüentemente para teste dos efeitos fixos). Nos dois casos, surge a questão do que significa uma amostra grande com dados desbalanceados. Ainda, segundo Henderson (1984), certamente N (o número total de dados) tendendo ao infinito não é uma condição suficiente para se ter uma grande amostra. Considerações a respeito do número de níveis em cada um dos efeitos aleatórios e da proporção de subclasses perdidas são importantes para responder essa questão. Conseqüentemente, a validade de uma aproximação qui-quadrado para o LRT e do próprio uso de uma χ^2 (para os quadrados médios) é incerta.

Henderson Júnior e Henderson (1979) consideraram quatro alternativas computacionais para a realização de testes de hipóteses com dados desbalanceados: (i) tratar os efeitos aleatórios como fixos; (ii) ignorar os efeitos aleatórios; (iii) tratar alguns efeitos aleatórios como fixos e ignorar outros; (iv) usar estimativas das variâncias dos efeitos aleatórios nas equações de modelo misto. Para a maioria das situações, tais autores preferem a alternativa (iv). O principal argumento baseia-se na lógica de que, no contexto dos modelos mistos, quando os efeitos aleatórios são tratados como fixos, tais modelos são tratados como se as suas variâncias (dos efeitos aleatórios) fossem de magnitudes muito altas em relação à variância do erro. Quando os efeitos aleatórios são ignorados, eles são tratados como se tivessem variância nula. A realidade, provavelmente, compreende uma situação intermediária entre estes dois extremos.

Do que foi exposto, torna-se claro o problema do teste dos efeitos fixos quando os componentes de variância são estimados a partir de dados desbalanceados. Mais recentemente, Berk (1987) propôs uma estatística para o teste simultâneo de vários fatores de efeitos fixos. Esta é denominada estatística de Wald (Wald, 1943), a qual é uma generalização da estatística T^2 de Hotelling (similar multivariado do teste t de Student) sem a necessidade de a matriz de variância-covariância seguir uma distribuição Wishart. A estatística W de Wald é dada por $W = (\hat{\beta} - \beta_o)' \hat{\Im} (\hat{\beta} - \beta_o)$ em que β_o é o vetor paramétrico de efeitos fixos sob $H_o: \beta = \beta_o$, $\hat{\beta}$ é a estimativa de β_o e $\hat{\Im}$ é a estimativa da matriz de informação de Fisher avaliada em $\hat{\beta}$.

Assintoticamente, a estatística de Wald tem distribuição χ^2, ou seja, $(\hat{\beta} - \beta_o) \hat{\Im} (\hat{\beta} - \beta_o) \sim \chi^2$. Assim, as estatísticas LRT e W são assintoticamente equivalentes e, sob H_o, convergem em distribuição para uma variável χ^2, sendo, entretanto, o LRT o critério que define um teste uniformemente mais poderoso (Silvey, 1975).

Os diferentes contrastes de efeitos fixos têm variâncias desiguais, cujas aproximações à distribuição χ^2 têm diferentes números de graus de liberdade. Assim, os graus de liberdade do denominador (erro) não são facilmente determinados na análise de um modelo misto desbalanceado. Segundo Gilmour et al. (2002), somente em poucos casos bem definidos, podem ser, precisamente, determinados os graus de liberdade do erro. A estatística de Wald distribuída como uma χ^2 considera que a variância do erro é conhecida, ou seja, assume infinitos graus de liberdade, sendo por isto mais frouxa.

Em resumo, na prática, três estatísticas têm sido recomendadas para teste dos efeitos fixos: (i) LRT (Welham e Thompson, 1997); (ii) W de Wald (Kenward e Roger, 1997; Elston, 1998); (iii) F de Snedecor (Gilmour et al., 2002). Segundo pesquisas recentes, a estatística de Wald para pequenas amostras é aproximada por uma distribuição F como uma razão entre as variáveis aleatórias χ^2. Assim, embora outras estatísticas possam ser introduzidas para o teste de efeitos fixos, a estatística de Wald é atrativa porque ela reproduz exatamente a análise de variância para delineamentos balanceados.

No software ASREML (Gilmour et al., 2002), os testes de hipóteses referentes aos efeitos fixos no contexto da estimação por REML têm sido baseados em tabelas de análise de variância (contemplando apenas os efeitos fixos), usando a própria estatística F de Snedecor. Em grandes análises, por exemplo, em modelos com grande número de efeitos aleatórios, os números de graus de liberdade serão altos e os testes estatísticos não serão muito sensíveis (Kennedy, 1991) a pequenas diferenças nos graus de liberdade do denominador. Segundo Gilmour et al. (2002), é preferível o teste F, em que se pode ter ao menos uma indicação (baseada na estrutura dos dados) de qual o número de graus de liberdade deve ser usado. A estatística de Wald superestima a significância (P valor menor do que deveria ser) quando o número de graus de liberdade do denominador do teste F não é muito grande. O LRT de Welham e Thompson é uma alternativa conservadora, pois subestima a significância do efeito, ou seja, superestima o P valor.

No contexto do melhoramento genético, pode haver o interesse em testar os efeitos fixos de blocos, por exemplo, visto que a diferença significativa entre blocos é desejável. Sendo significativo o efeito de blocos, os materiais genéticos são melhor testados para condições diversas de ambiente, permitindo a seleção de materiais mais estáveis, ou seja, ocorre uma redução nos efeitos da interação genótipo x ambiente em relação à seleção baseada no comportamento em vários blocos similares.

11.2 Teste de Efeitos Aleatórios

O uso da tabela de análise de variância para a construção de testes F para os efeitos aleatórios em modelos desbalanceados é muito difícil. Isto porque é necessária a obtenção dos

quadrados médios a partir dos componentes de variância e seus multiplicadores, que são muito difíceis de ser computados sob desbalanceamento.

Existe uma maneira mais formal para testar os efeitos aleatórios, ou seja, para verificar se determinado efeito aleatório necessita permanecer no modelo. Essa abordagem formal baseia-se no teste LRT. Um procedimento prático é descrito por Gilmour et al. (2002). Tendo-se a estimativa do componente de variância ($\hat{\sigma}^2$) do referido efeito aleatório e seu respectivo desvio padrão ($\hat{s}(\hat{\sigma}^2)$), pode-se concluir:

(i) se $\hat{\sigma}^2 - 2(\hat{s}\ (\hat{\sigma}^2)) > 0$ (ou alternativamente $\hat{\sigma}^2/(s\ (\hat{\sigma}^2)) > 1$), o efeito aleatório é significativo e deve permanecer no modelo. Essa assertiva é válida e conservadora pois:

- se $\hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2)) > 2$: o LRT será significativo.
- se $\hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2)) < 0{,}5$: o LRT será não significativo.
- se $1 < \hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2)) < 2$: existe a necessidade de aplicar o LRT.

A expressão $\hat{\sigma}^2 - 2(\hat{s}\ (\hat{\sigma}^2)) \rangle 0$ indica, com confiança maior do que 95 %, que o referido componente de variância não apresenta valor paramétrico igual a zero.

(ii) se $\hat{\sigma}^2 - 2\hat{s}\ (\hat{\sigma}^2) < 0$ ($ou\ \hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2)) < 1$), o efeito aleatório é não significativo e pode ser retirado do modelo.

Para ser um pouco mais exato, deve-se calcular $\hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2))$ e aplicar o LRT quando $1 < \hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2)) < 2$. A aplicação do LRT é apresentada com mais detalhes no tópico 10. Tal teste envolve duas vezes a redução no Log L resultante da retirada de t termos aleatórios, quantidade esta distribuída como uma χ^2_t. Assim, para a verificação da significância de um efeito aleatório, tem-se que LRT ~ χ^2_1. Entretanto, Stram e Lee (1994) sugerem uma correção por meio da multiplicação do P valor associado a χ^2_1 por 0,5, ou seja, sugerem o uso de uma distribuição $\chi^2_{0,5}$. Esta correção é, sobretudo, indicada para teste no limite do espaço paramétrico, quando o P valor aproximado

para a estatística de teste d (duas vezes a redução no Log L) é $0{,}5(1-P(\chi_1^2 \leq d))$, em que P denota probabilidade. Nesse caso (mistura de distribuições com 1 e 0 graus de liberdade), o valor tabelado de qui-quadrado para o nível de significância de 5 % é 2.79. Uma boa discussão sobre o assunto é apresentada por Crianiceanu e Ruppert (2004).

Dentre as três abordagens utilizadas para teste das significâncias de fatores de efeitos aleatórios (teste F, LRT ou análise de deviance, proporção $\hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2))$), o teste F é o mais acurado nas circunstâncias em que ele é válido, ou seja, na condição de balanceamento. Isto porque o mesmo faz pleno uso de toda a informação disponível. A significância baseada na análise de deviance não depende dos graus de liberdade associados ao fator em teste e, portanto, não considera completamente a consequência de um possível tamanho de amostra limitado. Mas, para o caso desbalanceado a análise de deviance é mais acurada que o teste F. A proporção $\hat{\sigma}^2/(\hat{s}\ (\hat{\sigma}^2))$ ou outra medida que utiliza $\hat{s}\ (\hat{\sigma}^2)$, não é muito confiável. Isto é devido à não normalidade da distribuição dos componentes de variância estimados. Conseqüentemente, o desvio padrão de uma variância não é tão informativo quanto aquele associado a um componente de média.

É importante enfatizar que os Log L não são comparáveis quando mudam-se os efeitos fixos ou a parte fixa do modelo.

12 REML PARA AVALIAÇÃO DE TRATAMENTOS DE EFEITOS FIXOS

REML/GLS, REML/BLUE ou simplesmente REML refere-se ao procedimento geral de análise de variância e estimação de efeitos de tratamentos, considerados fixos. É um procedimento análogo e generalizado em relação à ANOVA. Porém, difere no procedimento de estimação e é adequado a dados desbalanceados (perdas de parcela, representação desigual dos tratamentos, etc) e modelos mais complexos. O procedimento de estimação empregado é o da máxima verossimilhança residual (REML), o qual tende a produzir melhores estimativas do que o procedimento de quadrados mínimos empregado na ANOVA. As estimativas dos efeitos de tratamentos obtidas enquadram-se na classe BLUE (melhores estimativas lineares não viciadas).

Os modelos sujeitos à análise REML/GLS caracterizam-se pela definição dos efeitos de tratamentos como fixos e demais efeitos (a exceção da média geral) como aleatórios, podendo também haver outros efeitos fixos. Um modelo típico é aquele com tratamentos de efeitos fixos e blocos e resíduos como efeitos aleatórios. No caso, o procedimento REML/GLS produzirá estimativas (via REML) de componentes de variância associadas as fontes de variação blocos e resíduo e estimativas de efeitos de tratamentos pelo método de quadrados mínimos generalizados (GLS). De maneira genérica, o REML é um procedimento de estimação de componentes de variância e efeitos fixos. Em outras palavras, a função de verossimilhança residual a ser maximizada envolve componentes de variância e efeitos fixos.

Atualmente, o procedimento REML/GLS de Patterson e Thompson (1971) está substituindo o método ANOVA de Fisher (1925). O grande avanço computacional ocorrido a partir de 1990 e o surgimento de novos e eficientes algoritmos REML incluídos em softwares de excelência como o ASREML, GENSTAT e SAS contribuíram para isto (Littel, 2003). Após a estimação via REML, os testes de hipóteses e comparação de tratamentos podem ser realizados à semelhança do que é realizado tradicionalmente em associação com a ANOVA e com a estatística experimental tradicional.

13 MÉTODOS MATEMÁTICOS NUMÉRICOS PARA MODELAGEM E INFERÊNCIA VEROSSIMILHANÇA

O desenvolvimento de programas computacionais para análises via REML depende fortemente de conhecimentos de Cálculo Numérico e Álgebra Linear Numérica. Trata-se, portanto, de uma área essencialmente matemática. A resolução de sistema de equações lineares para obtenção do BLUP depende de métodos iterativos tais quais os de Jacobi, Gauss-Seidel, Gradientes Conjugados e Gradientes Conjugados Precondicionados. Esses três últimos são mais adequados para as aplicações em modelos lineares mistos. Essa área da Estatística é essencialmente fundamentada em Álgebra Linear (Searle, 1997; Harville, 1997).

Dada a importância do método numérico de Gauss-Seidel para a resolução iterativa de sistemas de equações lineares, este é descrito a seguir empregando um pequeno exemplo.

Seja o sistema de equações lineares:

$$\begin{cases} 4X_1 + X_2 + X_3 = 5 \\ -2X_1 + 5X_2 + X_3 = 0 \\ 3X_1 + X_2 + 6X_3 = -6,5 \end{cases}$$

As soluções para as três incógnitas X_1, X_2 e X_3 são dadas por:

$$X_1^k = \frac{(5 - X_2^{k-1} - X_3^{k-1})}{4}; \quad X_2^k = \frac{(0 + 2X_1^k - X_3^{k-1})}{5}; \quad X_3^k = \frac{(-6,5 - 3X_1^k - X_2^k)}{6},$$

em que k, refere-se à k-ésima iteração.

Partindo-se de um vetor inicial X^0 = (0, 0, 0), tem-se a 1ª iteração: $X_1^1 = \frac{(5-0-0)}{4} = \frac{5}{4}; \quad X_2^1 = \frac{(0+2\cdot 5/4-0)}{5} = \frac{1}{2}; \quad X_3^1 = \frac{(-6,5-3\cdot 5/4-1/2)}{6} = -1,7967.$

Na 2ª iteração, tem-se:

$$X_1^2 = \frac{(5-1/2+1,7967)}{4} = 1,58; \quad X_2^2 = \frac{(0+2\cdot 1,58-(-1,7967))}{5} = 0,992; \quad X_3^2 = \frac{(-6,5-3\cdot 1,58-0,992)}{6} = 2,03$$

O procedimento prossegue até que o menor valor de $\left| X^k - X^{k-1} \right| \leq \varepsilon$, em que ε é o erro desejado (geralmente $\leq 10^{-5}$).

Para maximização de funções de verossimilhança, ou seja, para obtenção do ponto de máximo de funções lineares ou não lineares de uma ou várias variáveis, os métodos numéricos são essenciais. Técnicas clássicas como os métodos de Newton-Raphson e de Gauss-Newton têm sido as principais ferramentas para esse propósito, com grande aplicação na área de Cálculo Matemático. Uma literatura básica inicial sobre o assunto é o livro de Ruggiero e Lopes (1996), adotado nos cursos de graduação em Matemática e Estatística. Os métodos de maior interesse da Estatística são apresentados a seguir.

13.1 Algoritmos Matemáticos Numéricos para Maximização de Função de Verossimilhança

Em geral, as equações de máxima verossimilhança ou vetor escore não têm solução explícita. Dessa forma, métodos matemáticos numéricos devem ser usados na maximização de verossimilhanças. Esse processo pode ser considerado como um problema matemático de otimização não linear. Os principais métodos usados são os de Newton-Raphson (NR), dos Escores de Fisher (FS), Esperança-Maximização (EM), livre de derivadas (DF) e de Informação Média (AI), este último usando uma média entre IOF e IEF. Adicionalmente, os métodos baseados em cadeias de Markov e método Monte Carlo (MCMC), muito usados em inferência bayesiana, podem também ser usados no contexto da inferência verossimilhança.

Os métodos empregados são geralmente iterativos. Em processos iterativos, inicia-se com certos valores atribuídos aos parâmetros da função, e iterações são realizadas por meio da atualização dos parâmetros até a obtenção da convergência, ou seja, até a obtenção de estimativas constantes dos parâmetros ou do valor máximo da função de verossimilhança. Nenhum dos métodos iterativos garante a convergência para o ponto de máximo global mas isto pode ser verificado reiniciando o processo com outros valores atribuídos aos parâmetros.

13.1.1 Método de Newton-Raphson (NR)

O algoritmo NR é um procedimento genérico para resolver uma função g(x) = 0. É um método padrão para obtenção de solução numérica de equações não lineares. Inicia-se com um valor x^0 e em seguida lineariza-se g(x) em torno de x^0 e posteriormente iguala-se a zero de forma a obter

$$g(x) \approx g(x^0) + g'(x^0)(x - x^0) = 0$$

A solução da equação linear propicia uma fórmula atualizada

$$x^1 = x^0 - g(x^0)/g'(x^0)$$

Para a estimação por máxima verossimilhança, o objetivo é resolver a equação escore $S(\theta) = 0$. Inicia-se com θ^0 e a fórmula de atualização é dada por $\theta^1 = \theta^0 - [S'(\theta^0)]^{-1} S(\theta^0) = \theta^0 + [I(\theta^0)]^{-1} S(\theta^0)$.

Genericamente, para a iteração (k + 1), tem-se:

$$\theta^{(k+1)} = \theta^k + [I(\theta^k)]^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k).$$

A diferença $\theta^{(k+1)}$ - θ^k é denominada ***correção*** e o processo iterativo continua até que a correção seja nula.

Verifica-se que a aproximação linear da função escore é equivalente à aproximação quadrática da função Log L. A convergência é quadrática e portanto bastante rápida.

O método NR deriva-se de uma Expansão em Série de Taylor para obter soluções de equações polinomiais e não lineares. Assim, tem-se a expansão $g(\theta) = g(\theta^0) + (1/1!)g'(\theta^0)(\theta - \theta^0) + (1/2!)g''(\theta^0)(\theta - \theta^0)^2 + (1/3!)g'''(\theta^0)(\theta - \theta^0)^3 ...$

Ignorando-se termos com graus elevados na expansão acima obtém-se aproximações quadráticas do tipo $g(\theta) = g(\theta^0) + g'(\theta^0)(\theta - \theta^0) + (1/2)g''(\theta^0)(\theta - \theta^0)^2$ e lineares do tipo $g(\theta) = g(\theta^0) + g'(\theta^0)(\theta - \theta^0)$.

Para a estimação REML sob modelos mistos, tem-se:

$\theta^{(k+1)} = \theta^k + [I(\theta^k)]^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k) = \theta^k - [H^k)]^{-1} \frac{\partial}{\partial \theta^k} \log L(\theta^k)$, em que H é a matriz Hessiano (matriz das derivadas segundas parciais da função Log L com respeito a todos os componentes de variância).

O método NR considera todos os termos lineares e quadráticos na expansão multidimensional em série de Taylor da função Log L com respeito aos componentes de variância. Têm-se as seguintes quantidades na equação acima:

$$H_{ij}^k = \frac{\partial^2}{\partial \sigma_i^2 \partial \sigma_j^2} LogL = 0.5 tr(P^{(k)} V_i P^{(k)} V_j) - y' P^{(k)} V_i P^{(k)} V_j P^{(k)} y$$

$$\frac{\partial}{\partial \theta^k} LogL(\theta^k) = \frac{\partial}{\partial \sigma_i^2} LogL(\sigma_i^2) \; = -0.5\, tr(P^{(k)} V_i) + 0.5\, y' P^{(k)} V_i P^{(k)} y$$

A matriz Hessiano para os componentes de variância aditiva e residual é dada por

$$H = \frac{\partial^2}{\partial \sigma_i^2 \partial \sigma_j^2} Log\, L = 0.5 \begin{bmatrix} tr(PP) - 2y'PPPy & tr(PA^*P) - 2y'PA^*PPy \\ tr(PA^*P) - 2y'PA^*PPy & tr(PA^*PA^*) - y'PA^*PA^*Py \end{bmatrix}$$, em que A* = ZAZ´.

Os traços no algoritmo NR envolvem, em forma complexa, elementos da matriz inversa das equações de modelo misto. Assim, são de difícil cômputo.

13.1.2 Método de Escores de Fisher (FS)

O método de escores de Fisher usa a matriz de informação esperada (IEF) em vez da matriz de informação observada (IOF), ou seja, usa o valor esperado da matriz Hessiano em vez de usar a matriz de informação observada de todas as derivadas parciais de segunda ordem.

O método FS requer menos cálculos porque muitas expressões matemáticas são eliminadas no processo de tomar as esperanças, ou seja, de se obter o valor esperado. Adicionalmente, a inversa da matriz de informação esperada fornece os desvios padrões das estimativas dos parâmetros. O método também é mais robusto (em relação ao NR) aos valores iniciais usados. No entanto, pode convergir em taxas mais lentas. Tanto o método NR quanto FS podem não convergir, especialmente quando a amostra de dados é pequena e o número de parâmetros é grande.

Genericamente, para a iteração (k + 1), tem-se:

$$\theta^{(k+1)} = \theta^k + [\Gamma(\theta^k)]^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k) = \theta^k + [E_\theta I(\theta^k)]^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k).$$

Por trabalhar com a informação esperada de Fisher, a matriz de informação esperada de Fisher ($F(\theta^k)$) substitui a matriz de informação observada do método NR. Ou seja, requer os valores esperados e não reais das derivadas segundas. Tem-se então:

$$\theta^{(k+1)} = \theta^k + [\Gamma(\theta^k)]^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k) = \theta^k + F(\theta^k)^{-1} \frac{\partial}{\partial \theta^k} LogL(\theta^k)$$

E a matriz $F(\theta^k)$ é dada por

$$F(\theta^k)_{ij} = 0.5tr(P^{(k)}V_iP^{(k)}V_j).$$

A matriz de informação esperada de Fisher para os componentes de variância aditiva e residual é dada por

$$F = -E(\frac{\partial^2}{\partial\sigma_i^2\partial\sigma_j^2}Log\ L) = 0.5\begin{bmatrix} tr(PP) & tr(PA*P) \\ tr(PA*P) & tr(PA*PA*) \end{bmatrix}.$$

A seguir, descrevem-se aspectos da aplicação do método FS na estimação de componentes de variância no melhoramento genético.

A função Log L pode ser rescrita em termos de proporcionalidade como $LogL \quad \alpha \quad y'Py - \log|V| - \log|X'V^{-1}X|$. A estimação do componente de variância θ_i envolve igualar a zero a derivada primeira (maximização), dada por $\frac{\partial}{\partial\theta_i}LogL = y'P(\frac{\partial V}{\partial\theta_i})Py - tr[P(\frac{\partial V}{\partial\theta_i})]$. Isto é análogo a igualar uma função dos dados à sua esperança matemática. Patterson e Thompson (1971) propuseram o uso do método FS, o qual tem o termo $E(\frac{\partial^2}{\partial^2\theta_i}LogL) = -0.5\,tr[P(\frac{\partial V}{\partial\theta_i})P(\frac{\partial V}{\partial\theta_i})]$ associado ao valor esperado da derivada segunda, o qual é a informação esperada (IE). Os valores de θ podem então ser atualizados usando $\hat{\theta} = \theta + IE^{-1}\frac{\partial}{\partial\theta}LogL$. Todos os termos dessa expressão advém da solução das equações de modelo misto e da inversa (C^{-1}) da matriz dos coeficientes desse sistema de equações.

Visando eliminar a necessidade de C^{-1} completa, Thompson (1977) propôs o cômputo de b por GLS e o cômputo de *a* como $\hat{a} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{b})$ de forma que a parte necessária de C^{-1} é obtida de $Var(\hat{a} - a) = (Z'R^{-1}Z + G^{-1})^{-1} + (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}X(X'V^{-1}X)^{-}X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}$, onde o segundo termo refere-se à correção para incerteza na predição *a*. O traço dessa correção contribui para o primeiro diferencial e, aplicando-se a regra de rotação, tal traço pode ser escrito como $tr[(X'V^{-1}X)^{-}X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}X]$. Isto demonstra que nem todos os elementos de C^{-1} necessitam ser calculados para formar o primeiro diferencial.

Algoritmos FS foram usados também por Meyer (1983; 1985). Em termos matriciais, o método FS sob modelo com dois componentes de variância pode ser aplicado resolvendo-se iterativamente o seguinte sistema matricial:

$$\begin{bmatrix} q-\lambda\ tr(A^{-1}C)+\lambda^2\ tr(A^{-1}C)^2 & tr(A^{-1}C)-tr(A^{-1}C)^2 \\ tr(A^{-1}C)-tr(A^{-1}C)^2 & NDF/\lambda^2+tr(A^{-1}C)^2 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_a^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \hat{a}'A^{-1}\hat{a} \\ \hat{e}'\hat{e}/\lambda^2 \end{bmatrix}$$, em que NDF = N – r(X) – q refere-se aos graus de liberdade do resíduo, q é o número de níveis no vetor a e C é a inversa da matriz dos coeficientes das equações de modelo misto referentes aos efeitos aleatórios após absorver b. Também $\lambda = \sigma_e^2/\sigma_a^2$, ou seja, equivale à relação entre as variâncias genética aditiva e residual.

A matriz dos coeficientes do sistema acima é proporcional à matriz de informação para σ_a^2 e σ_e^2, de forma que a inversa dessa matriz propicia estimativas de suas variâncias de amostragem.

13.1.3 Método Esperança - Maximização (EM)

O EM é um método para obtenção de estimativas de máxima verossimilhança e também de modas *a posteriori*. É um algoritmo estável numericamente significando que a cada iteração ocorre um aumento da verossimilhança ou densidade *a posteriori*, quase sempre conduzindo a convergência. Esse algoritmo propicia uma visão da estrutura estatística de um problema de máxima verossimilhança, ao contrário dos métodos NR e FS que são mais matemáticos por natureza, baseando-se fortemente em cálculo numérico.

O conceito do método refere-se à estimação em problemas denominados de dados incompletos. Entretanto, muitos problemas estatísticos que a primeira vista parecem não envolver dados perdidos podem também ser reformulados em termos de dados incompletos, por meio da técnica de aumento de dados com os valores não observados. Nesse caso, assume-se que parte dos dados foi observada (as próprios valores observados) e parte foi perdida (os parâmetros do modelo, a serem estimados).

Na especificação geral do EM, considera-se y como os dados incompletos e x como os dados completos. Tem-se que y é completamente determinado por x, ou seja, y = f(x). A recíproca

não é verdadeira. A idéia é que alguma informação se perde no caminho entre x e y e esta perda de informação está relacionada com a informação de Fisher.

Sendo y o conjunto de dados, o problema estatístico é estimar θ a partir da verossimilhança baseada em y, ou seja, $L(\theta; y) = p_\theta(y)$. Uma vez que a dependência em y é explícita, isto é distinto de $L(\theta; x) = p_\theta(x)$, que é a verossimilhança baseada em x.

O método EM obtém as estimativas de máxima verossimilhança $\hat{\theta}$ por meio do seguinte sistema iterativo a partir de um valor inicial θ^0:

(i) Passo E ou de cômputo da esperança ou valor esperado condicional de θ (processo de integração)

$Q(\theta) = Q(\theta|\theta^0) = E[\log L(\theta; x)|y, \theta^0]$. Esse passo calcula a esperança do Log L dos dados completos com base na distribuição condicional dos dados perdidos (parâmetros), com os dados observados e o corrente valor de θ.

(ii) Passo M ou de maximização de $Q(\theta)$ com respeito a θ para fornecer um valor atualizado θ^1. Esse valor é usado então no passo E e deve-se interagir nos dois passos até a convergência.

Se $\theta^{(t+1)}$ maximiza $Q(\theta|\theta^t)$, o passo M é tal que $Q(\theta^{(t+1)}|\theta^t) \geq Q(\theta|\theta^t)$, fato que implica que $\theta^{(t+1)}$ é uma solução para a equação $\partial Q(\theta|\theta^t)/\partial Q = 0$. A seqüência iterativa leva a um aumento monotônico de $Log\ p(\theta|y)$, conforme demonstrado por $Log\ p(\theta^{(t+1)}|y) \geq Log\ p(\theta^{(t)}|y)$. Esta é uma importante propriedade e torna o EM um procedimento numericamente estável a medida que ele escala a superfície de verossimilhança. Nesse aspecto, nenhuma garantia existe para o procedimento NR. Tal propriedade, entretanto, não garante convergência.

Outra vantagem prática do EM é que ele restringe as estimativas ao espaço paramétrico automaticamente. Isto porque cada passo M produz uma estimativa do tipo ML. As desvantagens em relação ao NR são:

- a convergência pode ser lenta, sendo que a velocidade de convergência é dependente da quantidade de informação perdida (número de parâmetros). Entretanto, existem procedimentos de aceleração da convergência que tornam o EM um algoritmo rápido (ver os tópicos 13.1.6 e 14). É possível manipular os dados completos x para minimizar a quantidade de informação perdida, por meio da expansão dos parâmetros.
- não produzem imediatos erros padrões das estimativas. Entretanto, se existe uma função Log L explícita, pode-se obter numericamente a informação de Fisher (IF) e então obter os erros padrões como o inverso da IF.

Existe uma similaridade entre os algoritmos NR e EM. Com o NR, obtém-se uma aproximação quadrática da função objetivo f(θ) em torno da estimativa inicial θ^0: $q(\theta) = f(\theta^0) + f'(\theta^0)(\theta - \theta^0) + (1/2) f''(\theta^0)(\theta - \theta^0)^2$ e encontra-se θ^1 como o maximizador de $q(\theta)$. O algoritmo converge rapidamente se f(θ) é bem aproximado por $q(\theta)$. Com o algoritmo EM, a função objetivo $L(\theta; y)$ é aproximada por $Q(\theta)$. Uma melhor aproximação de $L(\theta; y)$ por $Q(\theta)$ implicaria uma convergência mais rápida do EM.

No contexto dos modelos mistos, o algoritmo EM é dado conforme a seqüência.

Considerando o modelo $y = Xb + Za + e$, o passo E do algoritmo EM consiste em definir a função objetivo Q. Os parâmetros a e b são considerados dados faltantes e os dados observados y (denominados incompletos) são aumentados por esses parâmetros para fornecer os dados completos x. Então, os dados completos são naturalmente definidos por x = (y', b', a')'. O vetor b de efeitos fixos pode ser tratado como um vetor de efeitos aleatórios com variância infinita com o objetivo de caracterizar uma verossimilhança residual e b será eliminado pela integração na obtenção do REML. O vetor dos parâmetros de interesse (componentes de variância) é $\theta = (\sigma_a^2, \sigma_e^2)'$.

O passo E consiste em tomar os valores esperados do Log L dos dados completos, $L(\theta; x) = p(x|\theta)$, com respeito à distribuição condicional do vetor dos dados faltantes z = (b', a')', dado os dados observados e o vetor θ fixado com valor corrente $\theta^{(t))}$, ou seja, $Q = Q(\theta|\theta^t) = \int L(\theta; y; z)\, p(z|y, \theta = \theta^t) dz$.

Por definição, tem-se $L(\theta;x) = p(x|\theta) = p(y|b,a,\theta)\,p(b,a|\theta)$ em que

$p(y|b,a,\theta) = p(y|b,a,\sigma_e^2) = p(e|\sigma_e^2)$ e $p(b,a|\theta)\ \alpha\ p(a|\sigma_a^2)$. Assim, considerando a propriedade de separabilidade do logaritmo da verossimilhança, tem-se: $L(\theta;x) = L(\sigma_e^2;e) + L(\sigma_a^2;a) + const.$. Portanto, a função objetivo é dada por: $Q = Q(\theta|\theta^t) = Q_e(\sigma_e^2|\theta^t) + Q_a(\sigma_a^2|\theta^t) + const.$

A composição dessa função objetivo é dada por:

$Q_e(\sigma_e^2|\theta^t) = E_c^t[L(\sigma_e^2;e)] = -0.5[N\log 2\pi + N\log\sigma_e^2 + E(e'e|y,\theta^t)/\sigma_e^2]$ e

$Q_a(\sigma_a^2|\theta^t) = E_c^t[L(\sigma_a^2;a)] = -0.5[q\log 2\pi + q\log\sigma_a^2 + E(a'A^{-1}a|y,\theta^t)]/\sigma_a^2$, em que E_c^t denota esperança condicional.

A fase M envolve a maximização das funções $Q_e(\sigma_e^2|\theta^t)$ e $Q_a(\sigma_a^2|\theta^t)$ em função de σ_e^2 e σ_a^2, respectivamente. As derivadas primeiras dessas funções são dadas por:

$$\frac{\partial(-2Q_e)}{\partial\sigma_e^2} = \frac{N}{\sigma_e^2} - \frac{E_c^t(e'e)}{\sigma_e^4}$$ e

$$\frac{\partial(-2Q_a)}{\partial\sigma_a^2} = \frac{q}{\sigma_a^2} - \frac{E_c^t(a'a)}{\sigma_a^4}.$$

Igualando-se a zero essas derivadas, têm-se as seguintes fórmulas iterativas para os componentes de variância:

$$\sigma_e^{2(t+1)} = \frac{E_c^t(e'e)}{N}$$ e

$$\sigma_a^{2(t+1)} = \frac{E_c^t(a'A^{-1}a)}{q}.$$

Dessa forma, tais fórmulas são dependentes das esperanças condicionais.

Usando a regra da esperança matemática de produtos quadráticos (Searle, 1982) dada por $E(x'Mx) = \mu'M\mu + tr(MV)$, em que x é um vetor de variáveis aleatórias com média u e matriz de variância V e M é uma matriz quadrada, tem-se, considerando o parentesco entre os efeitos *a*:

$E(a'A^{-1}a|y,\hat{b},A) = \hat{a}'A^{-1}\hat{a} + tr(A^{-1}C^{22})$, no caso em que não se fatora a matriz dos erros (R), em ambos os lados do sistema de equações de modelo misto e $E(a'A^{-1}a|y,\hat{b},A) = \hat{a}'A^{-1}\hat{a} + \hat{\sigma}_e^2 tr(A^{-1}C^{22})$, no caso em que se fatora a matriz dos erros (R), em ambos os lados do sistema de equações de modelo misto. Nesse caso, $Var(a-\hat{a}) = PEV = \sigma_e^2 C^{22}$ e portanto $[E(a'A^{-1}a|y,\hat{b},A)]/q = \hat{\sigma}_a^2 = [\hat{a}'A^{-1}\hat{a} + \hat{\sigma}_e^2 tr(A^{-1}C^{22})]/q$ e $\hat{\sigma}_a^2 = [\hat{a}'A^{-1}\hat{a} + \hat{\sigma}_e^2 tr(A^{-1}C^{22})]/q$.

A esperança $E_c^t(e'e)$ é encontrada de maneira similar e é dada por $E(e'e|y,\hat{b},A) = \hat{e}'\hat{e} + tr[Var(\hat{e})]$, em que $\hat{e} = y - X\hat{b} - Z\hat{a}$. A fórmula iterativa para a variância residual é dada então por $\hat{\sigma}_e^{2(t+1)} = \{\hat{e}'\hat{e} + \hat{\sigma}_e^{2(t)}[tr(X) + q - \hat{\sigma}_e^{2(t)}/\hat{\sigma}_a^{2(t)} tr(A^{-1}C^{22})]\}/N$. Essa expressão difere da apresentada por Henderson (1973), dada por $\hat{\sigma}_e^2 = [y'y - y'X\hat{b} - y'Z\hat{a}]/[N - r(X)]$. Essa última expressão enquadra-se na categoria ECM ou EM condicional e é também muito eficiente computacionalmente.

Então para encontrar as esperanças, necessitam-se apenas da predição dos efeitos do modelo (b, a) e da obtenção da PEV. Essas estimativas são obtidas resolvendo-se as equações de modelo misto de Henderson.

Vários algoritmos EM desse tipo foram apresentados por Resende (2002) para diversos modelos. No entanto, para o modelo individual reduzido contemplando diferentes tipos de parentes não se obtém expressões explícitas como essas. Detalhes sobre isso são apresentados no tópico 14.

Em suma, o algoritmo EM pode ser resumido da seguinte forma.

Foi visto que $E(a'A^{-1}a|y,\hat{b}) = \hat{a}'A^{-1}\hat{a} + tr(A^{-1}PEV_a) = \hat{a}'A^{-1}\hat{a} + tr(A^{-1}C^{22})$.

Considerando o fundamento do cálculo de uma variância, tem-se $\hat{\sigma}_a^2 = [E(a'a|y,\hat{b})]/q = [\hat{a}'A^{-1}\hat{a} + tr(A^{-1}PEV_a)]/q = [\hat{a}'A^{-1}\hat{a} + tr(A^{-1}C^{22})]/q$ e, portanto,

$\hat{\sigma}_a^2 = [\hat{a}'A^{-1}\hat{a} + tr(A^{-1}C^{22})]/q$.

Para o método ML, as fórmulas iterativas são dadas por

$\hat{\sigma}_a^2 = [\hat{a}' A^{-1} \hat{a} + \hat{\sigma}_e^2 tr(Z' R^{-1} Z + A^{-1}\lambda)^{-1}] / q$ e

$\hat{\sigma}_e^2 = [y'y - y' X\hat{b} - y' Z\hat{a}] / N$.

Notam-se diferenças entre as expressões para o REML e ML. O ML não demanda inversão completa da matriz dos coeficientes e não corrige os graus de liberdade do resíduo para o número de efeitos fixos estimados.

No caso do método EM, os valores de θ são atualizados usando $\hat{\theta} = \theta + IO^{-1} \frac{\partial}{\partial \theta} LogL$, em que IO refere-se à informação observada obtida dos dados completos. Isto contrasta com $\hat{\theta} = \theta + IE^{-1} \frac{\partial}{\partial \theta} LogL$, associado ao método FS.

13.1.4 Método Livre de Derivadas (DF)

Por esse método, a função de verossimilhança residual pode ser avaliada sem a necessidade de obtenção das soluções das equações de modelo misto, sem a inversão da matriz dos coeficientes do sistema de equações e sem computar qualquer componente de variância isoladamente. Entretanto, várias avaliações são necessárias visando confirmar se a convergência foi para um máximo local ou global. Quanto mais parâmetros tiver o modelo, maior é a probabilidade de que a convergência ocorra para um máximo local. Assim, tal método tem limitações para analisar modelos com muitos componentes de variância.

Modelo incluindo um só efeito aleatório (aditivo-*a*)

O algoritmo DF-REML para modelos com apenas um fator aleatório além do resíduo, proposto por Graser et al. (1987), determina o ponto de máximo da função *Log L* por meio de sucessivas avaliações dessa função para valores de q (razão entre os componentes de variância

genética aditiva e residual). Assim, esse algoritmo não envolve a derivação da função *Log L* em relação aos componentes de variância, para a determinação dos estimadores dos componentes de variância. A seguir, será descrito o procedimento DFREML, para o modelo de plantas individuais com apenas um efeito aleatório.

Para um valor *a priori* de $q = \sigma_a^2 / \sigma_e^2$, pode-se estimar:

$\hat{\sigma}_e^2 = \frac{y'Py}{N - r(x)} = \frac{y'(y - x\hat{b} - Z\hat{a})}{N - r(x)}$, em que:

$P = V^{-1} - V^{-1} X (X'V^{-1}X)^{-} X'V^{-1}$: matriz de projeção ortogonal da parte aleatória das observações (*y*) no espaço coluna da matriz *X*.

$V = ZAZ'\hat{\sigma}_a^2 + I\hat{\sigma}_e^2 = Var(y)$

Para estimar $\hat{q} = \hat{\sigma}_a^2 / \hat{\sigma}_e^2$, são calculadas apenas as partes relevantes do máximo (q fixo) de *Log L* para vários valores de *q*, ou seja, *Log L* é maximizado com respeito apenas ao parâmetro q. Para avaliar a máxima verossimilhança para qualquer q, deve-se, também, avaliar $\hat{\sigma}_e^2$ e $\hat{\sigma}_a^2 = q\hat{\sigma}_e^2$. Neste caso, a função *Log L* a ser avaliada equivale a:

$LogL = -\frac{1}{2}\left[(N - r(x) - N_a)\log_e \hat{\sigma}_e^2 + \log_e |C| + N_a \log_e \hat{\sigma}_a^2 + y'Py / \hat{\sigma}_e^2\right]$, em que:

$|C|$: determinante de uma submatriz não singular de *C* com posto máximo, em que *C* é a matriz dos coeficientes das equações de modelo misto.

N_a: número de níveis no efeito aleatório *a*.

Uma vez que $\hat{\sigma}_e^2$ tenha sido calculado usando um valor *a priori* de q, a estimativa REML de $\hat{q}$ pode ser obtida diretamente da equação acima. Para avaliação de *Log L* é necessária a obtenção apenas de *y´Py e* $\log_e|C|$, uma vez que os demais termos são fixados.

Como a inversão de *V* (em *P*) demanda grande esforço computacional, Graser et al. (1987) propuseram uma estratégia computacional que baseia-se na absorção de *C* em *y´y* de forma a se obter $y'\left[V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}\right]y = y'Py$. A absorção de *C* em *y´y* não requer inversão, ou seja, é realizada pelo processo de eliminação de Gauss, absorvendo uma linha por vez, conforme Smith e

Graser (1986). Este método apresenta a vantagem adicional de se obter $log_e|C|$ pelo somatório do logaritmo dos elementos da diagonal C, ou seja, por meio dos pivôs não zero obtidos durante o processo de absorção via eliminação de Gauss. As computações são realizadas por meio da computação da seguinte matriz de absorção.

$$\begin{bmatrix} X'X & X'Z & X'y \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'y \\ y'X & y'Z & y'y \end{bmatrix} = \begin{bmatrix} C & & X'y \\ & & Z'y \\ y'X & y'Z & y'y \end{bmatrix} .$$

Assim, a inversão de C é evitada, por meio de sua expansão usando o vetor à direita das equações de modelo misto.

Na prática, três valores de q, em torno de seu valor esperado, devem ser utilizados. Uma função quadrática (equação de regressão, usando os três valores de *q*) pode ser calculada visando descrever *Log L* como função de *q*, e o ponto de máximo (valor de *q* que maximiza *Log L*) da equação quadrática fornece um novo valor de q. A seguir, deve-se reavaliar *Log L* com este novo valor de q e recalcular a equação quadrática usando este novo valor de q e mais dois adjacentes dentre os três anteriores. O procedimento deve ser repetido até a convergência para o valor de herdabilidade $[h^2 = q/(1+q)]$ na segunda casa decimal (Graser et al., 1987).

Modelo incluindo efeito aditivo (*a*) e um efeito aleatório adicional (*c*)

Para modelos que envolvem dois fatores aleatórios (além do resíduo) e três componentes de variância a serem estimados, Meyer (1989b) sugere a avaliação da razão θ entre o componente de variância e a variância fenotípica (somatória dos três componentes de variância) e não o parâmetro *q* (que é dado em relação a variância ambiental). Dessa forma, no presente caso necessitam ser avaliados os parâmetros θ_a e θ_c, relativos a σ_a^2 e σ_c^2, respectivamente.

A função *Log L* a ser avaliada, no presente caso, equivale a:

$$Log\ L = -\frac{1}{2}\left[(N - r(x) - N_a - N_c)\log_e \hat{\sigma}_e^2 + \log_e|C| + N_a \log_e \hat{\sigma}_a^2 + \log_e|A| + N_c \log_e \sigma_c^2 + y'Py/\sigma_e^2\right]$$, em que:

N_c: número de níveis no efeito aleatório c.

Neste caso, *V* (em *P*) refere-se a todos os efeitos aleatórios. Também $log_e|A|$ não depende dos parâmetros a serem estimados, ou seja, essa quantidade é constante e não necessita ser calculada para maximizar *Log L.*

Modelo incluindo efeito aditivo (*a*), de dominância (*d*) e um efeito aleatório adicional (*c*)

Neste caso, necessitam ser avaliados os parâmetros θ_a, θ_d e θ_c, relativos a σ_a^2, σ_d^2 e σ_c^2, respectivamente. Considerando normalidade, para a determinação do máximo da função de verossimilhança ou equivalentemente, -2 vezes o seu logaritmo, basta avaliar a função:

$$-2\ \log\ L = const + y'Py/\sigma_e^2 + \log_e|C| + [(N - r(X) - 2N_a - N_c)]\ \log_e\ \sigma_e^2 +$$
$$+ N_a\ (\log_e\ \sigma_a^2 + \log_e\ \sigma_d^2) + N_c\ (\log_e\ \sigma_c^2) + \log_e|A| + \log_e|D|$$

Uma vez que $log_e|A|$ e $log_e|D|$ não dependem dos parâmetros a serem estimados, os mesmos não necessitam ser calculados com a finalidade de maximizar *Log L*.

13.1.5 Método de Informação Média (AI)

No método EM, os valores de θ são atualizados usando $\hat{\theta} = \theta + IO^{-1}\frac{\partial}{\partial\theta}LogL$ e no método FS tais valores são atualizados usando $\hat{\theta} = \theta + IE^{-1}\frac{\partial}{\partial\theta}LogL$, em que IO refere-se à informação observada obtida dos dados completos e IE refere-se à informação observada.

Johnson e Thompson (1995), Gilmour, Thompson e Cullis (1995) e Jensen et al. (1997) apresentaram o algoritmo de Informação Média (AI), o qual baseia-se no uso de uma matriz de informação alternativa. Visto que as matrizes de IO e IE são difícil computação (pois envolvem a segunda derivada), tais autores propuseram o uso da matriz de informação média, a qual contempla uma média das matrizes IO e IE. O cálculo da matriz AI é muito mais simples do que o cálculo de qualquer uma das duas (IO e IE) isoladamente. Isto porque, quando é feita a média das derivadas segunda observadas e esperadas, o termo envolvendo traços de produtos da matriz inversa, são cancelados, permanecendo uma expressão de simples computação.

A matriz de informação média entre a matriz Hessiano (H) e de Informação de Fisher (F) é:

$$AI = (-H + F)/2 = 0.5\begin{bmatrix} y'PPPy & y'PA^*PPy) \\ y'PA^*PPy) & y'PA^*PA^*Py \end{bmatrix}.$$

A AI pode também ser denotada como $AI(\frac{\partial^2}{\partial\theta_i\partial\theta_j}LogL) = -0.5\,y'P\frac{\partial V}{\partial\theta_i}P\frac{\partial V}{\partial\theta_j}Py$.

A fórmula de atualização das estimativas dos parâmetros é dada por

$$\theta^{(k+1)} = \theta^k + [AI(\theta^k)]^{-1}\frac{\partial}{\partial\theta^k}\log L(\theta^k).$$

O cálculo dos elementos da matriz de informação média requer termos tais quais V_iPy, também denotados por $\frac{\partial V}{\partial\theta_i}Py$ e $\frac{\partial V}{\partial\theta_j}Py$. Tais termos podem ser calculados como variáveis de trabalho ($y(\sigma_i^2) = V_iPy$), seguido por obtenção de produtos cruzados residuais.

As variáveis de trabalho para a variância aditiva e residual são dadas por $y(\sigma_a^2) = A^*Py = (1/\sigma_a^2)Z\hat{a}$ e $y(\sigma_e^2) = Py = (1/\sigma_e^2)\hat{e}$, em que $\hat{a}$ é obtida como solução das equações de modelo misto e $\hat{e} = y - X\hat{b} - Z\hat{a}$.

O cálculo dos elementos da matriz AI pode ser obtido da forma relatada a seguir. Considerando o elemento y´PPPy da matriz AI, o cálculo é dado por $y'PPPy = y(\sigma_e^2)'Py(\sigma_e^2)$, em que $Py(\sigma_e^2)$ é o vetor coluna de resíduos. Os demais elementos da matriz AI podem ser calculados da mesma maneira. Esta foi a abordagem apresentada por Johnson e Thompson (1995).

Outra alternativa para a construção da matriz AI foi apresentada por Gilmour et al. (1995), por meio do processo de eliminação de Gauss (ou absorção) na matriz M abaixo.

$$M = \begin{bmatrix} y'R^{-1}y & y'R^{-1}X & y'R^{-1}Z \\ X'R^{-1}y & X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}y & Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_a^2 \end{bmatrix}$$

Após a eliminação de Gauss, o primeiro elemento (M(1,1)) da matriz M equivalerá a y´Py.

Substituindo y pela variável de trabalho para σ_e^2, ou seja, $y(\sigma_e^2)$, tem-se a matriz M_e:

$$M_e = \begin{bmatrix} y(\sigma_e^2)'R^{-1}y(\sigma_e^2) & y(\sigma_e^2)'R^{-1}X & y(\sigma_e^2)'R^{-1}Z \\ X'R^{-1}y(\sigma_e^2) & X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}y(\sigma_e^2) & Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_a^2 \end{bmatrix}$$

Após a eliminação de Gauss em M_e, o elemento $M_e(1,1)$ será igual a y´PPPy da matriz AI.

Por outro lado, substituindo y pela variável de trabalho para σ_a^2, ou seja, $y(\sigma_a^2)$, tem-se a matriz M_a:

$$M_a = \begin{bmatrix} y(\sigma_a^2)'R^{-1}y(\sigma_a^2) & y(\sigma_a^2)'R^{-1}X & y(\sigma_a^2)'R^{-1}Z \\ X'R^{-1}y(\sigma_a^2) & X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}y(\sigma_a^2) & Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_a^2 \end{bmatrix}$$

Após a eliminação de Gauss em M_a, o elemento $M_a(1,1)$ será igual a y´PA*PA*Py da matriz AI.

Para obter o produto cruzado ou diagonal secundária da matriz AI, a matriz M_{ae} deve ser formada da seguinte forma:

$$M_{ae} = \begin{bmatrix} y(\sigma_a^2)'R^{-1}y(\sigma_e^2) & y(\sigma_a^2)'R^{-1}X & y(\sigma_a^2)'R^{-1}Z \\ X'R^{-1}y(\sigma_e^2) & X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}y(\sigma_e^2) & Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_a^2 \end{bmatrix}$$

Após a eliminação de Gauss em M_{ae}, o elemento $M_{ae}(1,1)$ será igual a y´PA*PPy da matriz AI.

De posse da matriz AI completa, necessita-se de $\frac{\partial}{\partial\theta} LogL(\theta)$ para se ter

$\theta^{(k+1)} = \theta^k + [AI(\theta^k)]^{-1} \frac{\partial}{\partial\theta^k} LogL(\theta^k)$ completo na fórmula iterativa. Nesse sentido, tem-se:

$\frac{\partial}{\partial\sigma_i^2} LogL = -0.5tr(PV_i) + 0.5\ y'PV_iPy$, em que o componente $y'PV_iPy$ pode ser computado a partir das equações para $y(\sigma_e^2)$ e $y(\sigma_a^2)$. As derivadas parciais de primeira ordem necessárias são então:

$$\frac{\partial}{\partial\sigma_a^2} LogL = -0.5tr(PA^*) + 0.5\ y'PA^*Py = -0.5\ \{(q/\sigma_a^2) - [tr(A^{-1}C^{aa})/\sigma_a^4] - (\hat{e}/\sigma_e^2)'(Z\hat{a}/\sigma_a^2)\}$$

$$\frac{\partial}{\partial\sigma_e^2} LogL = -0.5tr(P) + 0.5\ y'PPy = -0.5\ \{[(N - r(X))/\sigma_e^2] - [q - tr(A^{-1}C^{aa})/\sigma_a^4](1/\sigma_e^2) - (\hat{e}'\hat{e}/\sigma_e^4)\}.$$

Em resumo, o procedimento AI é assim constituído:

a) Equação de iteração a ser resolvida: $\theta^{(k+1)} = \theta^k + [AI(\theta^k)]^{-1} \frac{\partial}{\partial\theta^k} LogL(\theta^k)$.

b) Obtenção da matriz $AI(\theta^k)$

b.1) Cálculo das variáveis de trabalho $y(\sigma_e^2)$ e $y(\sigma_a^2)$, usando soluções das equações de modelo misto e valores iniciais dos componentes de variância.

b.2) Construção das matrizes M, M_a, M_e e M_{ae}, realização da eliminação de Gauss e obtenção dos elementos da matriz $AI(\theta^k)$.

c) Obtenção das derivadas primeira em $\frac{\partial}{\partial \theta^k} LogL(\theta^k)$.

c) Resolução da equação em (a).

Em cada iteração é necessário o computo do Log L para verificar a convergência do processo iterativo. Isso deve ser realizado usando a expressão:

$$LogL = -\frac{1}{2}\left[(N - r(x) - N_a)\log_e \hat{\sigma}_e^2 + \log_e |C| + \log_e |A| + N_a \log_e \hat{\sigma}_a^2 + y'Py / \hat{\sigma}_e^2\right]$$

Atualmente, o método estatístico mais aceito para análise de experimentos é o REML. E o método numérico mais eficiente é o AI. Nesse sentido, o cientista Robin Thompson desempenhou papel essencial na criação do REML, na criação do método AI e no desenvolvimento dos *softwares* ASREML e GENSTAT.

13.1.6 Método Esperança - Maximização com Parâmetros Estendidos (PX-EM)

Este método é o mais recente (Liu et al., 1998; Foulley e Van Dyk, 2000) e também o mais eficiente juntamente com o AI. Vários textos, publicados em francês, descrevem com detalhes o método (Foulley, 2003; Foulley et al., 2002a e b; San Cristobal et al., 2002). Esse método baseia-se na normalização dos efeitos aleatórios e aumenta muito a velocidade de convergência quando comparado ao EM tradicional. Atualmente é utilizado na implementação dos *softwares* Wombat (antigo DFREML), ASREML e Selegen-REML/BLUP. No ASREML e Wombat é usado em associação com o AI.

No método PX-EM, o vetor dos parâmetros (componentes de variância) no modelo completo é expandido para $\theta = (\sigma_a^2, \sigma_e^2, \alpha)'$ em que α é um parâmetro de trabalho. O parâmetro de trabalho é incluído porque o método EM original (EMO) imputa os dados faltantes (σ_a^2, σ_e^2) sob um modelo errado, isto é, assumindo que as estimativas de máxima verossimilhança desses parâmetros são iguais àquelas em determinada iteração. Em outras palavras, em cada iteração, os parâmetros correntes são tratados como se eles maximizassem a função Log L. Assim, no início do processo iterativo (longe do máximo), a esperança da verossimilhança completa é computada com erro.

O algoritmo PX-EM capitaliza a diferença entre o valor imputado $\alpha^{(t+1)}$ de α e seu valor de referência α_0 para fazer o ajustamento em θ. Esse ajustamento é dado por $\theta_X^{(t+1)} - \theta_{EM}^{(t+1)} \approx r_{\theta|\alpha}(\alpha^{(t+1)} - \alpha_0)$, em que os dois primeiros termos referem-se aos valores de θ na iteração (t + 1) sob os algoritmos X e EM, respectivamente. O coeficiente r refere-se ao fator de regressão ou correção. Esse ajustamento melhora a taxa de convergência em termos do número de iterações necessário para a convergência. Assim, a eficiência do PX-EM pode ser atribuída aos parâmetros extra em α, os quais capturam informações não utilizadas no EMO.

Pelo método EM original, tem-se o modelo $y = Xb + Za + e$, com estrutura de variância

$a \sim N(0,\ G)$ $\qquad$ $E(y) = Xb$

$e \sim N(0,\ R)$ $\qquad$ $Var(y) = V = ZGZ' + R$.

No caso bivariado, tem-se:

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}; \quad a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix};$$

$$G = A \otimes G_O; \quad R = I \otimes R_O;$$

$$G_O = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{21}} & \sigma^2_{a_2} \end{bmatrix}; \quad R_O = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{21}} & \sigma^2_{e_2} \end{bmatrix} \quad ou \quad R_O = \begin{bmatrix} \sigma^2_{e_1} & 0 \\ 0 & \sigma^2_{e_2} \end{bmatrix}, \ em\ que:$$

$\sigma_{a_{12}}$: covariância genética aditiva entre os caracteres 1 e 2;

$\sigma_{e_{12}}$: covariância residual entre os caracteres 1 e 2.

Assim, pelo algoritmo EM, Var(a) = A ⊗ G_o.

Pelo método PX-EM o modelo com os efeitos aleatórios escalonados é dado por $y = Xb + Z(I \otimes \alpha)a^* + e$, com estrutura de variância $Var(a^*) = A \otimes G_0^*$. Pela definição de a* como normalizado ou re-escalonado por α, tem-se $G_0^* = \alpha^{-1} G_0 (\alpha^{-1})'$. Por esse modelo, os elementos de G_0^* são estimados pelo método EMO, assumindo α = I. Adicionalmente α é também estimado em

cada iteração. Posteriormente, as estimativas de G_0^* e α são usadas na estimação dos parâmetros de interesse G_0, por meio da expressão $G_0 = \alpha G_0^* \alpha'$. As variâncias e covariâncias residuais são também estimadas por EMO, porém após obtenção dos resíduos por $\hat{e} = y - X\hat{b} - Z(\hat{\alpha} \otimes I)\hat{a}^*$, ou seja, considerando os valores correntes de α em cada iteração.

O desvio $r_{\theta|\alpha}(\alpha^{(t+1)} - \alpha_0)$ fornece uma medida do erro em assumir α_0 no EM. O ajuste das estimativas de G_0 por $\hat{\alpha}$ é uma forma de regressar as estimativas pela diferença entre $\hat{\alpha}$ e seu valor assumido como I no método EMO. A obtenção de $\hat{\alpha}$ requer a estimação de K^2 parâmetros adicionais, em que K refere-se ao número de caracteres avaliados em cada indivíduo. A matriz das estimativas dos parâmetros $\hat{\alpha}$ é então quadrada de ordem K x K.

A estimação de $\hat{\alpha}$ é dada pelo seguinte sistema matricial

$$F\hat{\alpha} = h$$

$$\sum_{m=1}^{K} \sum_{n=1}^{K} f_{kl,mn}^{t} \alpha_{m,n}^{(t+1)} = h_{kl}^{t} \text{ ,}$$

em que:

k,l = 1, 2,..., K.

$$f_{kl,mn} = tr[Z_k' R^{-1} Z_m (\hat{a}_n \hat{a}_l' + \sigma_e^2 C^{a_n a_l})]$$

$$h_{kl} = \hat{a}_l' Z_k' R^{-1} y - tr[Z_k' R^{-1} X (\hat{b}\hat{a}_l' + \sigma_e^2 C^{ba_l})] .$$

Nas equações acima, C e seus sobrescritos referem-se a blocos da inversa da matriz dos coeficientes das equações de modelo misto, correspondentes às associações entre os efeitos descritos nos sobrescritos.

Para o caso univariado com um só efeito aleatório, α é um só parâmetro (desvio padrão do efeito aleatório) e o modelo para o algoritmo PX-EM é dado por $y = Xb + Z\alpha a^* + e$, em que $a^* = a/\alpha = a/\sigma_{a_\alpha}$. Nesse caso, a variância genética aditiva é atualizada em cada iteração por

$\sigma_a^{2(t+1)} = \alpha^2 \sigma_a^{2(t)} = (\sigma_{a_\alpha})^2 \sigma_a^{2(t)}$. Aqui, e com $R = I\sigma_e^2$, a estimação de $\hat{\alpha}$ é dada pelo seguinte sistema matricial $\hat{\alpha} = f / h$ em que:

$$f = tr[(\hat{a}'\hat{a}' + \sigma_e^2 C^{aa}]$$

$$h = \hat{a}' Zy - tr[Z' X (\hat{b}\hat{a}' + \sigma_e^2 C^{ba}]\,.$$

13.1.7 Métodos de Cadeias de Markov e Monte Carlo (MCMC)

O método estatístico REML e os métodos numéricos (NR, FS, EM, DF, AI e PX-EM) até aqui apresentados são denominados métodos exatos. Esses métodos são exatos no sentido de que não são baseados em amostragens de distribuições de probabilidade. Os métodos estatísticos bayesianos baseiam-se em amostragem e, nesse sentido, não são denominados métodos exatos. Os métodos numéricos empregados na abordagem bayesiana como a amostragem de Gibbs pertencem a uma classe de métodos denominada cadeias de Markov e Monte Carlo (MCMC). No entanto, para usar os métodos MCMC, não há necessidade de se empregar os fundamentos bayesianos. O fundamento dos métodos MCMC é de que, devido às dificuldades para se calcular as PEV associadas aos efeitos dos fatores aleatórios, essas são substituídas por amostragens. Assim, podem ser usados também associados ao algoritmo EM. Segundo Thompson (2002), nem sempre é claro qual abordagem computacional é mais eficiente: exata, amostragem de Gibbs bayesiana ou algo intermediário.

Segundo Schaeffer (1999), a amostragem de Gibbs é muito similar ao método iterativo de Gauss-Seidel, exceto que quando cada solução para os efeitos são obtidas, adiciona-se uma quantidade aleatória baseada na distribuição condicional *a posteriori* de sua variância. Para usar a amostragem de Gibbs, há necessidade apenas de um programa de resolução das equações de modelo misto, um bom gerador de números aleatórios e tempo computacional para processar um imenso número de amostras. Thompson (2002) relata um procedimento de aumento de dados para reduzir o esforço computacional na estimação de componentes de variância, porém sem adicionar tanto “noise” em *a*. O procedimento envolve o ajuste de dois modelos $y - Z\tilde{a} = Xb + e$ e

$y - X\tilde{b} = Za + e$. No primeiro modelo ajusta-se $\hat{b}$ e se obtém $\tilde{b} = \hat{b} + amostragem$. No segundo modelo, ajusta-se y para $\tilde{b}$, estima-se σ_a^2 e σ_e^2, ajusta-se $\hat{a}$ e obtém-se $\tilde{a} = \hat{a} + amostragem$. Então ajusta-se y para $Z\tilde{a}$ e o procedimento é repetido. Após um período de aquecimento, as médias $\overline{\sigma}_a^2$ e $\overline{\sigma}_e^2$ fornecem estimativas para σ_a^2 e σ_e^2, assim como no procedimento de amostragem de Gibbs. Isto evita adicionar tanto noise em $\tilde{a}$ quando σ_a^2 e σ_e^2 são estimados.

Também Harville (2004) propõe o uso da amostragem de Gibbs como uma forma de tornar o REML computacionalmente possível para grande conjuntos de dados e modelos complexos.

13.1.8 Método Esperança - Maximização Estocástico (SAEM)

O algoritmo esperança – maximização com aproximação estocástica (SAEM) foi apresentado por Jaffrezic et al. (2007) como uma forma eficiente de computação e inferência em modelos não lineares mistos. Nessa situação complexa, geralmente são usados procedimentos aproximados de máxima verossimilhança e também métodos bayesianos. O método SAEM surge como uma opção de rápida convergência em relação aos algoritmos EM Monte Carlo e bayesiano. Outra vantagem é que o mesmo não requer a especificação de distribuições *a priori* e é bastante robusto à escolha dos valores iniciais no processo iterativo. A idéia é reciclar os valores simulados de uma iteração, na próxima iteração do algoritmo EM, fato que acelera consideravelmente a convergência. Uma proposta de aceleração da convergência do SAEM, via uso de parâmetros estendidos (PX), foi realizada por Lavielle e Meza (2007) por meio do algoritmo PX-SAEM.

14 ASPECTOS COMPUTACIONAIS PARA OBTENÇÃO DO REML

A obtenção de estimativas REML é muito mais complexa do que a obtenção de predições BLUP. Os algoritmos computacionais dependem de vários outros aspectos além da simples escolha do método numérico para maximização da função. A escolha de um procedimento computacional

adequado depende dos seguintes fatores: (i) escolha do modelo misto equivalente (individual ou reduzido) mais apropriado para a estrutura dos dados; (ii) escolha do método numérico mais atrativo; (iii) escolha dos algoritmos matriciais para lidar com matrizes esparsas ou densas e para fatoração da matriz dos coeficientes das equações de modelo misto, dependendo da estrutura dos dados e tipo de modelo a ser ajustado; (iv) escolha dos métodos de aceleração da convergência mais adequados às estruturas dos dados e tipo de modelo; (v) escolha da linguagem de programação mais adequada à programação matemática de grande complexidade; (vi) dimensionamento fixo ou dinâmico de vetores e matrizes no processo de fatoração; (vii) escolha da ordem de processamento das equações de modelo misto.

Dentre esses fatores, o de mais fácil escolha é a linguagem de programação matemática. As mais recomendadas são o Fortran e C^{++}. O dimensionamento fixo ou dinâmico de vetores e matrizes no processo de fatoração determina se: (i) haverá necessidade de compilação de novas versões do programa para cada tamanho do conjunto de dados e modelo a ser ajustado (dimensionamento fixo); (ii) o dimensionamento será automático após a leitura do modelo e do conjunto de dados (alocação dinâmica). O dimensionamento dinâmico é preferível, porém mais laborioso.

A escolha dos algoritmos matriciais quanto a esparsidade das matrizes depende da situação, e os principais métodos para cálculo da inversa de matrizes esparsas foram descritos por Takahashi et al. (1973), Zollenkof (1971) e George e Liu (1981). Esses métodos calculam somente os elementos da inversa que pertencem ao padrão de esparsidade da matriz original. Mesmo assim, o custo computacional para o cálculo da inversa esparsa é de duas a três vezes maior do que para cálculo de determinantes. O cálculo de uma inversa esparsa aumenta os requerimentos computacionais para avaliação de verossimilhanças. Thompson et al. (1994) apresentaram métodos para encontrar os elementos da matriz esparsa, os quais reduzem esses requerimentos.

Os métodos numéricos adequados à obtenção de estimativas REML foram apresentados nos tópicos anteriores. A implementação computacional da metodologia de modelos mistos baseia-se fortemente em métodos numéricos, notadamente, em álgebra linear numérica, visando à obtenção iterativa das soluções das equações de modelo misto (obtenção do BLUP) e, cálculo numérico para a maximização/minimização de funções de várias variáveis, visando à obtenção das estimativas REML.

Vários algoritmos computacionais para a obtenção de componentes de variância por ML e REML têm sido desenvolvidos tais como o FS, o EM (Dempster et al., 1977), o DF-REML (Graser et al., 1987) e o AI (Johnson e Thompson, 1995). Dentre estes, os mais usados são o EM e o AI-REML. O algoritmo EM é muito estável, numericamente, apresentando convergência mesmo que os valores iniciais não tenham sido totalmente adequados. Entretanto, uma inconveniência do algoritmo EM é a lentidão para as estimativas próximas ao limite do espaço paramétrico (por exemplo, quando uma variância tende a zero). Se valores iniciais positivos forem utilizados, a convergência para valores não negativos é garantida (Harville, 1977).

O algoritmo EM atua por meio da obtenção da esperança (por integração) e maximização (derivação) da função de verossimilhança dos dados, sucessivamente. Nos modelos de plantas individuais, em que, freqüentemente, a ordem das equações de modelo misto excedem o número de observações, a obtenção de estimativas por meio de primeira derivada pelo método EM requer a inversão da matriz dos coeficientes das equações de modelo misto, aumentando muito o esforço computacional. Segundo Lynch e Walsh (1997), os métodos de Newton-Raphson e de Fisher apresentam convergência quadrática, ao passo que o algoritmo EM apresenta convergência linear, sendo, portanto, mais lento.

Para contornar esta questão, Graser et al. (1987) propuseram um algoritmo para obtenção do ponto de máximo do logaritmo da função de verossimilhança por meio de sucessivas avaliações da função, partindo de valores atribuídos aos componentes de variância. Assim, o máximo relativo aos componentes de variância é encontrado por um processo de procura direta, sem requerer a inversão da matriz dos coeficientes. Por não envolver a derivação da função densidade de probabilidade, em relação aos componentes de variância, para o estabelecimento do sistema de equações a ser utilizado no processo iterativo, o algoritmo foi denominado DFREML. O algoritmo DF requer menos tempo de processamento por iteração do que o EM, porém exige maior número de iterações para atingir a convergência.

Recentemente, o algoritmo denominado AI-REML (*Average Information-REML*) foi desenvolvido e tem sido muito utilizado (Gilmour et al., 1995; Johnson e Thompson, 1995; Meyer, 1997). Segundo Johnson e Thompson (1995), o algoritmo DF apresenta propriedades numéricas pobres, e as soluções apresentam baixa acurácia nos dígitos significativos, o que constitui um

problema quando vários componentes de variância são estimados. Assim, o algoritmo AI é muito competitivo em relação aos algoritmos DF e EM e converge em menos de dez iterações, embora o tempo por iteração seja duas a três vezes maior que aquele requerido pelo algoritmo DF. Mesmo assim, o tempo total para convergência é muito menor. Segundo os últimos autores, para modelos complexos, o algoritmo AI é cinco vezes mais rápido que o DF e três vezes mais rápido que o algoritmo EM.

Os algoritmos para obtenção de estimativas REML podem ser agrupados de acordo com a ordem das derivadas usadas. Assim, têm-se: (i) não derivativo (DF-REML); (ii) baseado em derivadas parciais de primeira ordem (EM-REML); (iii) baseado em derivadas parciais de primeira e segunda ordens (AI-REML). O algoritmo AI é um procedimento derivativo melhorado, o qual fundamenta-se no uso dos métodos de Newton, que usam as derivadas primeira e segunda da função de verossimilhança. Tal algoritmo fundamenta-se na utilização da informação advinda da média das derivadas segundas observadas e esperadas da função de verossimilhança, de forma que o termo que contém os traços dos produtos da matriz inversa é cancelado, restando uma expressão mais simples para computação. Técnicas de matrizes esparsas são empregadas no cálculo dos elementos da inversa da matriz dos coeficientes, os quais são necessários para as derivadas primeiras da função de verossimilhança. Este algoritmo é também denominado Quasi-Newton (Gilmour et al., 1995), o qual aproxima a matriz hessiano (matriz de derivadas segundas) pela média das informações observadas e esperadas. A informação observada é uma medida da curvatura da função (ou do seu log) de verossimilhança e a informação esperada é a própria informação de Fisher.

Os algoritmos DF ganharam popularidade devido às suas flexibilidades quanto aos modelos (Meyer, 1989; 1991) e por causa da disponibilidade de *softwares*. Entretanto, as dificuldades de convergência em modelos mais complexos geraram um novo interesse em algoritmos baseados em primeira e segunda derivadas da função de verossimilhança, como o AI. Hofer (1998) mostrou que os algoritmos DF são vantajosos computacionalmente somente quando o número de parâmetros a serem estimados é pequeno. Os algoritmos EM são mais eficientes que o DF quando o número de parâmetros é grande. Nesta mesma situação, o algoritmo AI supera o EM.

Dentre os modelos mistos equivalentes, o modelo individual reduzido (MIR) permite a redução da dimensão do vetor de efeitos de um dos fatores aleatórios, no caso o vetor de efeitos genéticos. No entanto, o modelo individual (MI) é mais fácil de ser construído computacionalmente e, no caso do REML pelo método numérico EM, o MI permite a derivação de equações explícitas para os componentes de variância. Entretanto, com um número de indivíduos na casa dos milhares, dezenas de milhares e muitas vezes centenas de milhares de indivíduos a serem avaliados, o MI pode se tornar proibitivo.

A avaliação genética de campo via REML/BLUP contempla a matriz de parentesco completa e envolve entidades genéticas (indivíduos, linhagens, etc) com estruturas genealógicas variadas contemplando: (i) indivíduos que são genitores e portanto apresentam avaliações próprias e também em seus descendentes; (ii) indivíduos que não são genitores e portanto apresentam apenas avaliações próprias, e possuem os seus dois genitores (pai e mãe) conhecidos (por exemplo, indivíduos de progênies de irmãos completos); (iii) indivíduos que não são genitores e portanto apresentam apenas avaliações próprias, e possuem apenas um de seus dois genitores (pai ou mãe) conhecidos (por exemplo, indivíduos de progênies de meios irmãos).

Nesta situação, o componente de variância residual contempla três quantidades distintas: variação ambiental mais a variação não aditiva no caso dos indivíduos enquadrados na situação i; variação ambiental mais a variação não aditiva mais (1/2) da variação genética aditiva no caso dos indivíduos enquadrados na situação ii; variação ambiental mais a variação não aditiva mais (3/4) da variação genética aditiva no caso dos indivíduos enquadrados na situação iii. Assim, não há pelo método EM, uma fórmula explícita para isolar da variação residual associada ao MIR, os componentes ambiental e não aditivo, do componente aditivo. Situações idênticas são aquelas em que se tem somente um tipo de progênie (meios ou completos irmãos) e mais dados também dos genitores.

Considerando somente a predição BLUP, o MIR é muito vantajoso sobre o MI. Entretanto, para a estimação REML, vários pesquisadores acreditam que os benefícios computacionais propiciados pelo MIR são cancelados pela grande dificuldade em construir as equações (White et al., 2006). Henderson (1986) apresentou equações para a estimação de componentes de variância por EM. Essas equações envolvem a computação de formas quadráticas para os fatores aleatórios e sub-

equações para as variâncias dos erros de predição (PEV) dos efeitos de todos os fatores aleatórios. Tomando os traços dos produtos das formas quadráticas pelas PEV obtém-se p+1 equações para p parâmetros ou componentes de variância. Somando-se as duas equações referentes ao fator aleatório dos efeitos genéticos aditivos, o sistema de equações pode ser resolvido para os p componentes de variância associados aos p fatores aleatórios. A seguir maiores detalhes são apresentados.

No modelo individual reduzido, o modelo individual y = Xb + Za + e é desmembrado em:

Modelo individual reduzido para os genitores

$y_g = X_g b + Z_g a_g + e$

Modelo individual reduzido para os não genitores

$y_n = X_n b + Z_1 a_g + e^*$, em que:

X_n : matriz de incidência para efeitos fixos, referentes aos não genitores.

Z_1 : matriz de incidência, cujos únicos elementos não zero equivalem a ½, identificando os genitores dos indivíduos (o valor ½ advém da consideração de que $a - a_d = ½ (a_p + a_m)$).

a_g : vetor de valores genéticos dos genitores.

e^* : vetor de erros aleatórios : $e^* = a_d + e$.

Para os indivíduos não genitores, a_d e e podem ser combinados em um único resíduo, $e^* = a_d + e$.

Assim, o modelo individual reduzido – MIR – pode ser escrito como:

$$\begin{bmatrix} y_g \\ y_n \end{bmatrix} = \begin{bmatrix} X_g \\ X_n \end{bmatrix} b + \begin{bmatrix} Z_g \\ Z_1 \end{bmatrix} a_g + \begin{bmatrix} e \\ e^* \end{bmatrix}$$

Denominando-se:

$$X = \begin{bmatrix} X_g \\ X_n \end{bmatrix},\ W = \begin{bmatrix} Z_g \\ Z_1 \end{bmatrix}\ e\ R = \begin{bmatrix} R_g \\ R_n \end{bmatrix} = \begin{bmatrix} I\sigma_e^2 & 0 \\ 0 & I\sigma_{e^*}^2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I + D * \alpha^{-1} \end{bmatrix} \sigma_e^2$$

as estruturas de variância são:

$$Var(y) = ZA_g Z'\ \sigma_a^2 + R$$
$$Var(a_g) = A_g\ \sigma_a^2$$

em que:

A_g : é a matriz de parentesco para genitores e D* é uma matriz diagonal com elementos iguais a d_j, em que d_j = ½, ¾ ou 1, se ambos, um ou nenhum dos genitores são conhecidos, respectivamente.

$\sigma_{e^*}^2 = \sigma_e^2 + d_j\ \sigma_a^2 = (1 + d_j\ \alpha^{-1})\ \sigma_e^2$, ignorando a endogamia.

$\alpha = \sigma_e^2 / \sigma_a^2$

Assim, as equações de modelo misto são:

$$\begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + A_g^{-1}(1/\sigma_a^2) \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

A matriz R pode ser rescrita como $R = D\sigma_a^2 + R_e\sigma_e^2$ em que $R_e = I$.

Para a estimação dos componentes de variância são necessárias duas formas quadráticas para o vetor de erros preditos ($\hat{e}$) e uma forma quadrática para o vetor de valores genéticos preditos ($\hat{a}$). A forma quadrática para $\hat{a}$ é dada por $\hat{a}'Q\hat{a}$, sendo a matriz associada igual a $Q \equiv A_g^{-1}\sigma_a^{-4}$. As duas formas quadráticas para $\hat{e}$ são dadas por $\hat{e}'P_a\hat{e}$ e $\hat{e}'P_e\hat{e}$. As duas matrizes associadas são iguais a $P_a = R^{-1}DR^{-1}$ e $P_e = R^{-1}R_eR^{-1}$.

Essas formas quadráticas devem ser igualadas às suas esperanças matemáticas para que se obtenha equações resultantes para $\hat{\sigma}_a^2$ e $\hat{\sigma}_e^2$. Para encontrar essas esperanças deve-se observar que:

$E(\hat{a}'Q\hat{a}) = tr[Q\ \mathrm{var}(\hat{a})]$,

$E(\hat{e}'P_a\hat{e}) = tr[P_a\ \mathrm{var}(\hat{e})]$ e

$E(\hat{e}'P_e\hat{e}) = tr[P_e\ \mathrm{var}(\hat{e})]$.

Verifica-se assim que, para encontrar os valores esperados necessitam-se das PEV dos efeitos aleatórios, ou seja, Var($\hat{a}$) e Var($\hat{e}$) e essas são funções lineares de $\hat{\sigma}_a^2$ e $\hat{\sigma}_e^2$ e podem ser obtidas conforme descrito a seguir.

A inversa da matriz dos coeficientes das equações de modelo misto é da forma

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + A_g^{-1}(1/\sigma_a^2) \end{bmatrix}^{-1} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} = \begin{bmatrix} C^1 \\ C^2 \end{bmatrix} = C^{-1}.$$

Denominando W = [X : Z], tem-se então:

$$Var(\hat{a}) = (C^1W'R^{-1}ZA_gZ'R^{-1}WC^{1'} + C^1W'R^{-1}DR^{-1}WC^{1'})\sigma_a^2 + C^1W'R^{-1}R_eR^{-1}WC^{1'})\sigma_e^2 .$$

Sendo $\hat{e} = (I - WC^{-1}W'R^{-1})y = Ty$, então

$$Var(\hat{e}) = (T'ZA_gZ'T + T'DT)\sigma_a^2 + T'R_eT)\sigma_e^2 .$$

Finalmente, tomando traços em $Q[\mathrm{var}(\hat{a})]$, $P_a[\mathrm{var}(\hat{e})]$ e $P_e[\mathrm{var}(\hat{e})]$, obtém-se três equações em $\hat{\sigma}_a^2$ e $\hat{\sigma}_e^2$. Somando-se as duas primeiras, obtém-se a equação para $\hat{\sigma}_a^2$ e a terceira equação é referente a $\hat{\sigma}_e^2$. O processo é repetido iterativamente até a convergência.

White et al. (2006) reavivaram o MIR para a estimação de componentes de variância e propuseram algumas outras alternativas para a obtenção do REML. Tais alternativas envolvem a sobreposição de matrizes de incidência para os fatores aleatórios bem como o tratamento de valores perdidos nessas matrizes de incidência. O fato é que essas alternativas envolvem a criação adicional de equações para alguns indivíduos e não para outros. Adicionalmente, são impostas restrições aos componentes de variância, criando-se extra variâncias, em alguns casos, fixando-se valores negativos para essas extra variâncias. Assim, a escolha entre MI e MIR nem sempre é óbvia.

Alguns métodos numéricos para o REML, tais quais o EM apresentam convergência lenta. Assim, torna-se fundamental adotar algum procedimento de aceleração da convergência. Mantysaari e Van Vleck (1989) sugeriram uma aceleração ao algoritmo EM com base na taxa de convergência geométrica observada. Foulley e Quaas (1995) e Meng e Van Dyk (1998) sugeriram uma decomposição Cholesky ou parametrização linear como um meio de aumentar a velocidade de convergência do EM. Foulley e Quaas (1995) usam o modelo $y = Xb + \sigma_G Za^* + e$ e dado σ_G fazem a predição de a com as equações de modelo misto. Regressões de y em σ_G e Za* (levando em conta a incerteza de a) propicia uma atualização natural de σ_G (mantendo σ_G^2 no espaço paramétrico).

Também, Liu et al. (1998) propuseram o algoritmo parâmetros expandidos (PX) para acelerar o EM. Foulley e Van Dik (2000) compararam as taxas de convergência dos algoritmos PX-EM com o EM e ECME (ou EM condicional) e concluíram que o PX-EM geralmente converge mais rápido. Meyer (2006) relata que o algoritmo PX-EM possui eficiência comparável ao do AI de Gilmour et al. (2005) e Johnson e Thompson (2005). Adicionalmente, tal autora relata que resultados mais confiáveis e rápidos são obtidos com a combinação (PX-EM x AI) dos dois algoritmos, usando o PX-EM nas primeiras iterações para explorar a estabilidade e boa performance e então usar o AI para explorar a rápida convergência próxima ao Log L. Esse esquema híbrido foi sugerido por Cullis et al. (2004) e Thompson et al. (2005). Esses últimos autores relatam que os únicos inconvenientes do método AI são: (i) nem sempre melhora a verossimilhança a cada iteração (mas essa dificuldade se reduz na medida em que os parâmetros se aproximam do valor máximo da verossimilhança); (ii) podem conduzir a estimativas fora do espaço paramétrico. Esses inconvenientes são evitados quando se usa esse esquema híbrido.

Quanto ao ordenamento das equações de modelo misto, a escolha da ordem de processamento é fundamental para diminuir os requerimentos de memória e o tempo de processamento. Os algoritmos EM, AI e DF têm requerimentos de memória similares. Mas o AI envolve muito mais cálculos por iteração e muito menos iterações que os demais, pois apresenta convergência quadrática. Gilmour e Thompson (2006) relatam procedimentos ótimos de ordenamento baseados em algoritmos de listas encadeadas.

15 MODELOS LINEARES MISTOS GENERALIZADOS: REML PARA VARIÁVEIS NÃO NORMAIS

Variáveis não normais e não contínuas, como aquelas com distribuição binomial e outras variáveis categóricas, não são bem descritas por modelos estatísticos lineares. Para estas variáveis discretas, os modelos não lineares podem ser mais apropriados.

A classe de modelos lineares generalizados (MLG), desenvolvida por Nelder e Wedderburn (1972), permite a generalização ou flexibilização dos modelos lineares clássicos de variáveis contínuas, de forma que toda a estrutura para a estimação e predição em modelos lineares normais pode ser estendida para os modelos não lineares. Os modelos lineares clássicos são casos especiais de modelos lineares generalizados.

Estes modelos generalizados foram desenvolvidos para análise de dados associados a distribuições pertencentes à família exponencial com um parâmetro. Nelder e Wedderburn (1972) introduziram a idéia de modelos lineares generalizados, visando permitir maior flexibilidade de análise. Tal idéia relaxa a suposição de que *Y* segue distribuição normal e permite que esta siga qualquer distribuição que pertença à família exponencial na forma canônica. As generalizações ocorrem em duas direções: (i) permitem que a esperança μ, de *Y* seja uma função monotonicamente diferençável do preditor linear $\eta = \sum x_i\beta_i$ de forma que $\mu = f(\eta) = f(\sum x_i\beta_i)$; (ii) ou, por inversão, $g(\mu) = \eta$, em que g é a função de ligação, a qual liga a média ao preditor linear. A incorporação da função de ligação nas equações de modelos lineares mistos para e estimação de componentes de variância e de efeitos fixos e predição de variáveis aleatórias gera a denominação de modelo não linear devido à relação não linear que existe entre a escala latente e a probabilidade de um indivíduo pertencer a uma determinada categoria da variável discreta.

Para dados binomiais, $0 \leq \mu \leq 1$, funções de ligação tal qual a logito são utilizadas para satisfazer esta restrição natural. As transformações são importantes para: (i) estender a amplitude da variável analisada de (0,1) para a reta real; (ii) fazer a variância constante através da amplitude dos efeitos fixos (na escala da variável latente contínua). A função de ligação descreve, então, a relação existente entre o preditor linear (η) e o valor esperado μ de *Y* . No modelo linear clássico,

tem-se $\eta = \mu$ que é chamada ligação de identidade, e esta ligação é adequada no sentido em que ambos η e μ podem assumir valores na reta real (Mc Cullagh e Nelder, 1989).

As distribuições a serem assumidas para a escala da variável latente e correspondentes funções de ligação devem ser capazes de transformar o intervalo (0,1) em (- ∞, ∞). Neste sentido, as distribuições logística, normal padrão e Gumbel (ou distribuição de valor extremo) para a variável latente e suas correspondentes funções de ligação denominadas logito, probito e complemento log-log são apropriadas para o modelo binomial.

Para as situações em que Y segue uma distribuição binomial, a estimação em modelos lineares generalizados com efeitos fixos e aleatórios, pode ser realizada conforme Schall (1991).

A função de ligação logito, $g(\mu)$, aplicada aos dados y, é linearizada, conforme a expansão em série de Taylor de primeira ordem, fornecendo y*, da seguinte forma:

$$y^* = g\ (y) = g\ (\mu) + (y - \mu)\ g'(\mu)$$

Assim, tem-se:

$$y_i^* = \eta_i + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}$$

$$= \log\left(\frac{\mu_i}{1-\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}.$$

De posse da variável observacional ou dependente ajustada y*, tem-se que o modelo linear misto equivale a $y^* = Xb + Za + (y - \mu)g'(\mu)$, em que:

$$E(y^*) = Xb,\ Cov(a) = G,\ Cov[(y-\mu)g'(\mu)] = W^{-1}\sigma_e^2\ e\ Cov(y^*) = ZGZ' + W^{-1}\sigma_e^2.$$

W é uma matriz diagonal com elementos $w_i = (\partial\mu_i / \partial\eta_i)^2 / [\mu_i\ (1-\mu_i)]$.

O modelo $y^* = Xb + Za + (y - \mu)g'(\mu)$ tem a mesma estrutura da primeira e segunda ordem que o modelo $y = Xb + Za + e$, de forma que os algoritmos de estimação e predição para o caso normal podem ser adaptados, apenas, substituindo y por y* e $Cov(e) = R$ por $Cov[(y-\mu)g'(\mu)] = W^{-1}\sigma_e^2$.

Assim, têm-se as seguintes equações de modelo misto:

$\begin{bmatrix} X'S^{-1}X & X'S^{-1}Z \\ Z'S^{-1}X & Z'S^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b}_L \\ \hat{a}_L \end{bmatrix} = \begin{bmatrix} X'S^{-1}y^* \\ Z'S^{-1}y^* \end{bmatrix}$, em que:

S^{-1}: matriz com termos diagonais dados por $\mu_i\ (1-\mu_i)\frac{1}{\sigma^2_{e_L}}$;

$\sigma^2_{e_L}$: variância residual na escala contínua latente.

$b_L\ e\ a_L$ = efeitos fixos e aleatórios na escala latente.

Um modelo linear misto ponderado, conforme descrito acima, pode ser ajustado via REML. Os estimadores dos componentes de variância pelo método REML são dados por:

$$\hat{\sigma}^2_{a_L} = \frac{\hat{a}'_L\,\hat{a}_L}{q - tr\ C^{22}/\sigma^2_{a_L}}; \quad \hat{\sigma}^2_{e_L} = \frac{(y - X\hat{\beta}_L - Z\hat{a}_L)'\ S^{-1}(y - X\hat{\beta}_L - Z\hat{a}_L)}{N - r(x) - q + tr\ C^{22}/\sigma^2_{a_L}}\ \sigma^2_{e_L}$$

Quando *a* se refere a um vetor de valores genéticos aditivos, tem-se que $Cov(a) = G = A\sigma^2_{a_L}$, em que A é a matriz de correlação genética aditiva entre os indivíduos e $\sigma^2_{a_L}$ é a variância de a_L. Os estimadores REML são dados por:

$$\hat{\sigma}^2_{a_L} = \frac{\hat{a}'_L\,A^{-1}\,\hat{a}_L}{q - tr\ (A^{-1}\ C^{22})/\sigma^2_{a_L}}; \quad \hat{\sigma}^2_{e_L} = \frac{(y - X\hat{b}_L - Z\hat{a}_L)'\ S^{-1}(y - X\hat{b}_L - Z\hat{a}_L)}{N - r(x) - q + tr\ (A^{-1}\ C^{22})/\sigma^2_{a_L}}\ \sigma^2_{e_L}$$

O processo iterativo é repetido até a convergência, com o valor predito de $\hat{\theta} = X\hat{b}_L + Z\hat{a}_L$, transformado, usando a função de ligação para obtenção do novo valor predito de μ através de $\hat{\mu} = \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}$, o qual é utilizado para atualização de S^{-1} e y*.

Em resumo, o processo de estimação envolve:

a) Estimação de $\mu = n_1 / N$, em que n_1 é o número de indivíduos que recebem o escore 1, dentre N indivíduos avaliados.

b) Obtenção de y*, a partir de y e μ (neste passo, a variável passa do intervalo (0,1) para a reta real, ou seja, a função é linearizada).

c) Estimação de $\hat{b}_L \; e \; \hat{a}_L$, dados os valores atuais ou correntes de $\mu, \; \sigma^2_{e_L} \; e \; \sigma^2_{a_L}$.

d) Obtenção de $\hat{\sigma}^2_{e_L} \; e \; \hat{\sigma}^2_{a_L}$, iterativamente e, após a convergência, proceder à obtenção atualizada de $\hat{b}_L \; e \; \hat{a}_L$.

e) Obtenção de $\hat{\eta} = \hat{\theta} = X\hat{b}_L + Z\hat{a}_L$.

f) Obtenção de novo valor predito de μ, usando a função de ligação, por meio de $\hat{\mu}_L = \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}$ (neste passo, a variável volta ao intervalo (0,1)).

g) Atualização de S^{-1} via $S^{-1} = \hat{\mu}_1(1-\hat{\mu}_1)\frac{1}{\hat{\sigma}^2_{eL_1}}$ e de y* via $y^* = \log\left(\frac{\hat{\mu}_1}{1-\hat{\mu}_1}\right) + \frac{y-\hat{\mu}_1}{\hat{\mu}_1(1-\hat{\mu}_1)} = \hat{\theta} + \frac{y-\hat{\mu}_1}{\hat{\mu}_1(1-\hat{\mu}_1)}$.

h) Voltar ao passo (c), enquanto não atingir a convergência.

Note-se que este algoritmo é essencialmente hierárquico, havendo, a cada iteração, compreendendo os passos de (a) até (h), a necessidade de convergência no passo (d).

Os algoritmos apresentados são do tipo EM-REML. Outros algoritmos relatados em literatura para a análise de modelos lineares generalizados mistos (GLMM) são o PQL (*penalized quasi-likelihood*) de Breslow e Clayton (1993) e o IRREML (*iterated re-weighted* REML) de Engel e Keen (1994). Para GLMM's, os procedimentos de estimação e predição de Schall (1991), Breslow e Clayton (1993) e Engel e Keen (1994) são equivalentes (Keen e Engel, 1997). Sob suposições de normalidade, para componentes de variância fixos, o procedimento IRREML é equivalente ao procedimento bayesiano MAP (máximo *a posteriori*), de forma que o IRREML fornece uma alternativa, não bayesiana, de derivação do MAP (Engel e Keen, 1996).

O uso de modelos não lineares é especialmente indicado quando os indivíduos a serem avaliados pertencem a diferentes níveis dos efeitos fixos, com diferentes valores de incidência para a variável em análise. Neste caso, a obtenção do coeficiente de herdabilidade na escala base contínua, por meio da transformação de probito é inapropriada, pois a transformação é função da incidência, a qual difere para os indivíduos dos diferentes níveis dos efeitos fixos. Uma vez que as diferenças entre os

níveis dos efeitos fixos correspondem a mudanças na escala base, uma função para ligação das mudanças nas duas escalas necessita ser incorporada nas equações de modelo misto. Para cômputo de y* e S^{-1}, diferentes valores de μ_i devem ser computados para os diferentes níveis dos efeitos fixos. Se a média da variável não varia muito através dos níveis dos efeitos fixos, o uso da técnica GLMM não é necessária.

Maiores detalhes sobre a estimação e predição em modelos lineares generalizados mistos são apresentados por Mc Cullock e Searle (2001), Pawitan (2001), Resende (2002) e Resende e Biele (2002).

16 QUASE-VEROSSIMILHANÇA E EQUAÇÕES DE ESTIMAÇÃO GENERALIZADA (GEE) PARA ANÁLISE MULTIVARIADA DE VARIÁVEIS NÃO NORMAIS

Análises estatísticas univariadas de variáveis discretas são realizadas eficientemente via a classe de modelos lineares generalizados. Nesse caso, uma função de verossimilhança é maximizada iterativamente analisando uma variável linearizada (transformação de y para a escala linear), usando modelos lineares normais ponderados. Modelos mistos normais ponderados podem ser ajustados via REML.

Para o caso multivariado, a estatística clássica tem se limitado a técnicas descritivas não paramétricas tal qual a análise de componentes principais ou a modelos paramétricos baseados em normalidade. Em muitas aplicações, principalmente na área de estatística médica (Brown e Kempton, 1994; Brown e Prescott, 1999), muitos problemas de estimação associados a variáveis discretas não podem ser abordados usando a estatística multivariada tradicional. Para o caso de variáveis não normais, uma forma geral para a distribuição multivariada não existe. Isto conduz ao fato de que uma verdadeira função de verossimilhança, que baseia-se em normalidade, não está disponível. Uma função alternativa é a quase-verossimilhança, a qual tem propriedades similares às da verossimilhança verdadeira. Essa função de quase-verossimilhança pode ser maximizada usando a técnica das equações de estimação generalizada (GEE) criada por Liang e Zeger (1986), Zeger e Liang (1986) e Zeger et al. (1988). Por essa técnica, a estimação pode ser realizada por meio do método numérico ou algoritmo de quadrados mínimos ponderados iterativos (IWLS). Então, a técnica

GEE encontra seu principal uso na análise multivariada de variáveis discretas. É então um desdobramento da classe de modelos lineares generalizados (GLM) em que se incorporam as correlações entre variáveis ou entre medidas repetidas. Pode ser aplicada a modelos de efeitos fixos (Liang e Zeger, 1986) e a modelos de efeitos mistos (Zeger et al., 1988).

Uma diferença fundamental entre uma verossimilhança verdadeira e uma quase-verossimilhança abordada via equações de estimação é referente aos modelos de trabalho. Esses, no primeiro caso, tratam a verossimilhança como uma função objetivo para estimação e comparação de modelos. E no caso da quase-verossimilhança somente uma equação escore é especificada e resolvida para produzir uma estimativa. Essa abordagem da equação de estimação (EE) focaliza apenas o parâmetro de interesse e não toda a estrutura de probabilidade das observações. Uma vantagem da verossimilhança verdadeira refere-se à possibilidade de comparação de modelos via deviance e AIC. Uma abordagem alternativa de estimação associada à quase verossimilhança refere-se ao procedimento da pseudo-verossimilhança (Davidian e Giltinan, 1993), o qual permite a comparação de modelos via LRT e AIC.

A análise de modelos lineares generalizados pode ser gerada via equações de estimação, via pseudo verossimilhança ou via REML ou IWLS (abordagem de verossimilhança verdadeira), mas as filosofias subjacentes são diferentes. Uma distinção essencial é que o teste da razão de verossimilhança não está disponível na abordagem EE. De maneira genérica, as seguintes aplicações têm sido mais usadas.

Classe de Modelos	Dimensão do Modelo	Função Associada à Variável Aleatória Discreta	Classificação do Modelo quanto aos Efeitos	Método de Estimação	Algoritmo Numérico
Modelos Lineares Generalizados (GLM)	Univariada	Verossimilhança	Fixo	Máxima Verossimilhança (ML)	Quadrados Mínimos Ponderados Iterativos (IWLS)
Modelos Lineares Generalizados (GLM)	Multivariada	Quase-Verossimilhança	Fixo	Equações de Estimação Generalizada (GEE)	Quadrados Mínimos Ponderados Iterativos (IWLS)
Modelos Lineares Generalizados Mistos (GLMM)	Univariada	Verossimilhança Residual	Misto	Máxima Verossimilhança Residual (REML)	Vários
Modelos Lineares Generalizados Mistos (GLMM)	Multivariada	Quase-Verossimilhança	Misto	Pseudo Máxima Verossimilhança ou REML Condicional	Vários

Um exemplo trivial de equação de estimação é o método dos momentos. A EE aplicada ao caso multivariado é denominada equações de estimação generalizada (GEE). Por essa técnica e para a variável y, a estimativa de um parâmetro θ é dada por:

$\sum_{i=1}^{n}(\partial\mu_i/\partial\theta)\, v_i^{-1}(y_i-\mu_i)=0$, em que:

$E(y_i)=\mu_i(\theta)$; $\quad Var(y_i)=v_i(\theta)$.

Para o caso multivariado em um modelo de efeitos fixos, a GEE de Liang e Zeger (1986) é dada por:

$\sum_{i=1}(\partial\mu_i/\partial\theta)V_i^{-1}(y_i-\mu_i)=0$, em que y_i e v_i agora são vetores e V_i é uma matriz de variância, dada por $V_i=\sigma^2[H_i(\mu_i)]^{1/2}R_i(\gamma)[H_i(\mu_i)]^{1/2}$, onde:

$H_i(\mu_i)$: matriz diagonal de variâncias, de dimensão v x v, em que v é o número de variáveis ou de medições, e com elementos diagonais dados pela função de variância.

$R_i(\gamma)$: matriz de correlação, de dimensão v x v, que depende de um vetor de parâmetros γ, e, em seu caso mais simples $R_i(\gamma) = I$.

A função objetivo associada com a equação acima é denominada quase-verossimilhança e apresenta duas características marcantes (Pawitan, 2001):

(i) Em contraste com a verossimilhança completa ou verdadeira, nenhuma estrutura de probabilidade é especificada, mas somente as funções da média e variância. Assim, essa abordagem pode ser denominada semi-paramétrica, em que os demais parâmetros, exceto aqueles de interesse, são deixados livres. Especificando apenas a média e a variância, a forma da distribuição permanece totalmente livre.

(ii) Com essa modelagem limitada, a amplitude de inferências possíveis é também limitada. Basicamente, apenas uma estimativa pontual do parâmetro é obtida. A construção de intervalos de confiança e a realização de testes de hipóteses assumem normalidade assintótica das estimativas, produzindo uma inferência do tipo Wald. Também, a comparação de modelos é limitada.

17 MODELOS LINEARES MISTOS GENERALIZADOS HIERÁRQUICOS (HGLMM)

Nos modelos lineares mistos generalizados, descritos nos tópicos anteriores, assume-se que os residuos podem não apresentar distribuição normal, mas, os demais efeitos aleatórios do modelo seguem a distribuição normal. Entretanto, essa suposição nem sempre é adequada. Um exemplo é a situação em que os dados seguem distribuição de Poisson e a função de ligação especificada para os resíduos é a logarítmica. Nesse caso, uma suposição mais apropriada para os demais fatores aleatórios é uma distribuição gama com função de ligação logarítmica. Modelos em que uma distribuição de probabilidade e uma função de ligação podem ser especificados para cada fator aleatório são denominados modelos lineares mistos generalizados hierárquicos (HGLMM) (Lee e

Nelder, 1996; 2001). Como os fatores aleatórios nem sempre são de classificação hierárquica, uma denominação alternativa é modelos lineares mistos generalizados estratificados. HGLMM's são bem descritos por Lee et al. (2007).

CAPÍTULO 5

ANÁLISE ESTATÍSTICA ESPACIAL

1 ESTATÍSTICA DESCRITIVA ESPACIAL E VARIOGRAMA

Conjuntos de dados oriundos de pesquisas nas áreas de Ciências da Terra e do Meio Ambiente (Geologia, Hidrologia, Engenharia Florestal, etc) e Experimentação Vegetal possuem como atributo distinto determinada localização no espaço. Assim, informações referentes ao grau de continuidade e tendências espaciais das variáveis podem ser fundamentais na análise deste tipo de conjunto de dados. As técnicas de estatística descritiva apresentadas no Capítulo 2 não são adequadas para o estudo de características espaciais de um conjunto de dados. Novas estatísticas e técnicas descritivas são necessárias e vários livros de estatística espacial descrevem essas técnicas (Isaacks e Srivastava, 1975; Valente, 1990).

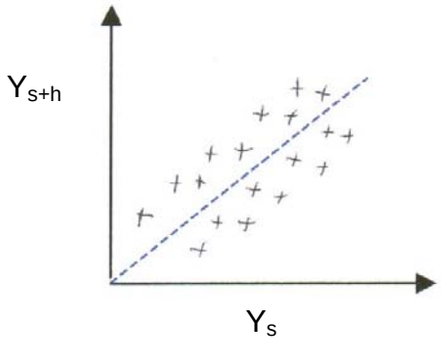
1.1 Diagramas Espaciais

As variáveis com comportamento espacial são denominadas variáveis regionalizadas e mostram características intermediárias entre as variáveis verdadeiramente casuais ou aleatórias e aquelas completamente determinísticas, exatas ou matemáticas. A estatística clássica trata de variáveis aleatórias ao passo que a estatística espacial aborda estas variáveis mistas.

Tais variáveis regionalizadas apresentam uma aparente continuidade no espaço. A continuidade geográfica se manifesta pela tendência de a variável apresentar valores muito próximos (dependentes) em dois pontos vizinhos e muito diferentes em pontos distantes. Assim, não são realizações de uma variável aleatória, pois são correlacionadas.

Para estudo destas variáveis, são essenciais os mapas ou croquis dos experimentos, mostrando a posição das observações bem como a vizinhança. Estes mapas e croquis, em conjunto com os dados experimentais, são utilizados na geração de diagramas ou gráficos de dispersão, estatísticas e gráficos bi ou tri-dimensionais, contemplando a variabilidade espacial dos experimentos. Estes gráficos são denominados variogramas.

Os diagramas de dispersão utilizados na descrição de duas variáveis X e Y podem também ser usados para descrever a relação entre o valor de uma variável e o valor desta mesma variável em pontos próximos. Neste caso, um diagrama mostra todos os possíveis pares de valores cujas posições são separadas por uma certa distância h em uma determinada direção. Assim, tem-se na abscissa (ou eixo x) os valores (y_s) da variável y na posição s e na ordenada (eixo y) os valores da variável ($y_{(s+h)}$), ou seja, a variável y na posição (s + h). Cada par (y_s, $y_{(s+h)}$) representa um ponto no diagrama, que tem a forma:

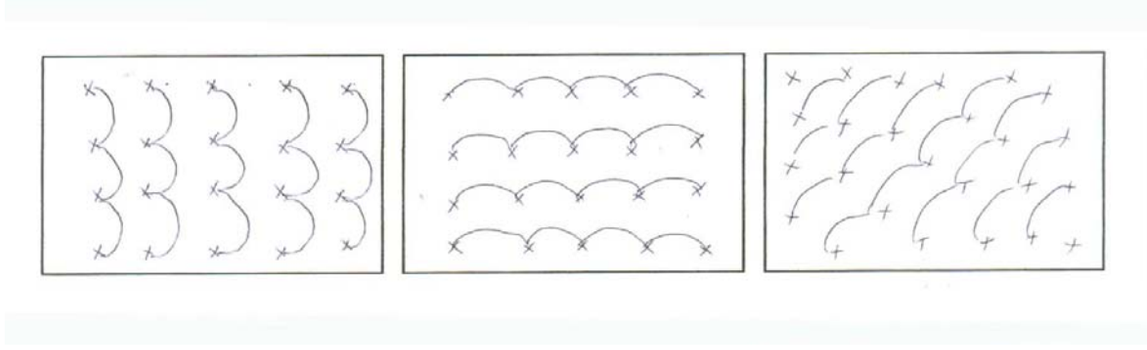


Os pares de pontos podem ser tomados em várias distâncias (1, 2, 3, ..., n metros) ou "*lags*" (vizinhança de primeira ordem, de segunda ordem, n) e em várias direções: (i) norte-sul ou sentido dentro de colunas na grade experimental (h = (0, n)); (ii) oeste-leste ou sentido dentro de linhas na grade experimental (h = (n, o)); (iii) diagonal, h = (n, n), por exemplo h = (1, 1). Ilustrações destes pares de pontos são apresentados a seguir:

(i)
Direção norte-sul ou dentro de colunas, distância 1 metro ou vizinhança de 1ª ordem (h = (0,1)). Total de 15 pares de pontos.

(ii)
Direção oeste-leste ou dentro de linhas, distância 1 metro ou vizinhança de 1ª ordem (h = (1,0)). Total de 16 pares de pontos.

(iii)
Direção diagonal ou nordeste, distância $\sqrt{2}$ (dada pela hipotenusa de um triângulo retângulo com espaçamento de 1 metro entre observações), vizinhança diagonal de 1ª ordem em forma de matriz Toeplitz h = (1, 1). Total de 16 pares de pontos.

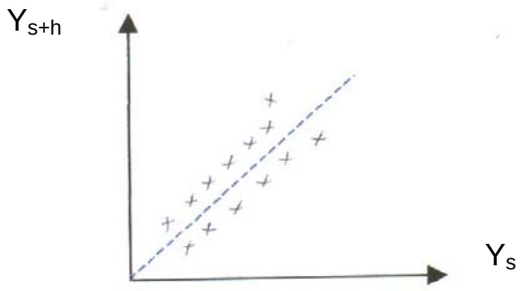


Estes pares de pontos são utilizados para a composição de diagramas y_s versus $y_{(s+h)}$. A forma da nuvem de pontos neste diagrama informa sobre a continuidade dos valores observados sobre determinada extensão em uma dada direção. Se os valores em posições separadas por uma

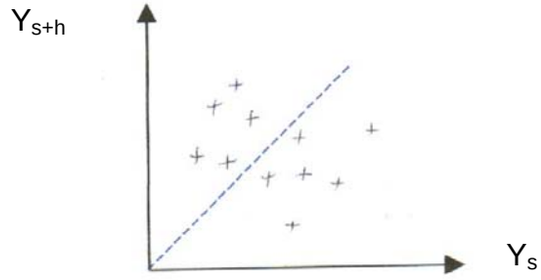
distância h são similares, tem-se que os pontos associados aos pares de observações serão plotados próximo à reta de 45^o passando pela origem, ou seja, ter-se-ia X = Y no plano cartesiano. Na medida em que os dados tornam-se menos similares, a nuvem de pontos torna-se mais grossa e difusa.

As figuras seguintes ilustram esta questão, para vizinhanças de primeira e quarta ordens.

Vizinhança de 1ª ordem (h = 0,1)



Vizinhança de 4ª ordem (h = 0,4)



Verifica-se que na vizinhança de maior ordem os pontos espalham-se mais, distanciando da reta de 45^o.

1.2 Correlograma, Variograma e Covariograma

Algumas estatísticas permitem sumarizar as informações contidas nos diagramas e descrever a continuidade espacial. Estas estatísticas são: (i) o coeficiente de correlação entre valores separados por uma distância h, ou seja, o coeficiente de autocorrelação, também denominado autocorrelação serial ou autocorrelação espacial; (ii) a covariância entre valores separados por uma distância h ou autocovariância; (iii) momento de inércia ou semivariância.

O coeficiente de autocorrelação diminui com o aumento da distância entre as observações, ou seja, informa que os pontos vão se distanciando da linha de 45^o. O estimador do coeficiente de autocorrelação é o mesmo do coeficiente de correlação linear simples entre duas variáveis.

Entretanto, em conformidade com a notação no contexto espacial, tal estimador no intervalo h é dado por:

$$\rho(h) = \frac{Cov\ (y_s, y_{(s+h)})}{[Var\ (y_s)\ Var\ (y_{(s+h)})]^{1/2}}$$

A relação entre o coeficiente de autocorrelação e a distância h é denominada correlograma ou função de autocorrelação. O coeficiente ρ(h) depende de h, o qual é um vetor e, portanto, possui uma magnitude e direção. Em geral, o correlograma ρ(h) versus h é plotado para cada uma das direções de interesse.

A autocovariância dada por $\gamma\ (h) = Cov\ (y_s, y_{(s+\ h)})$ também diminui com o aumento da distância entre as observações. A relação entre a autocovariância e a distância h é denominada função de autocovariância ou covariograma. O estimador da autocovariância no intervalo h é dado por:

$$\gamma(h) = Cov(y_s, y_{(s+h)}) = \frac{1}{N(h)} \sum y_s \cdot y_{(s+h)} - m_{y_s} \cdot m_{(y_{(s+h)})}$$, em que:

N(h): número de pares de dados separados por uma distância h;

m_{ys} e $m_{(y(s+h))}$: média dos dados y_s e $y_{(s+\ h)}$;

$$m_{y_s} = \frac{1}{N(h)} \sum y_s \ ;$$

$$m_{(y_{(s+h)})} = \frac{1}{N(h)} \sum y_{(s+h)} \ .$$

Outra expressão alternativa é:

$$\gamma(h) = \frac{1}{N(h)} \sum y_s \cdot y_{(s+h)} - m^2$$, em que m é a média geral de todos os dados.

Estas duas expressões são diferentes, uma vez que m_{ys} e $m_{y(s+h)}$ são diferentes na prática e portanto $m^2 \neq m_{ys} \ . \ m_{(y(s+h))}$. Portanto, a primeira expressão é preferida.

A autocorrelação ρ(h) refere-se à covariância padronizada pelos apropriados desvios padrões, sendo dada alternativamente por:

$$\rho(h) = \frac{\gamma(h)}{\sigma_{y_s}\ \sigma_{y_{(s+h)}}}$$, em que:

$$\sigma^2_{y_s} : \frac{1}{N(h)} \sum y_s - m^2_{y_s};$$

$$\sigma^2_{y_{(s+h)}} : \frac{1}{N(h)} \sum y_{(s+h)} - m^2_{y_{(s+h)}}$$

A semivariância ou momento de inércia em relação à reta de 45^o é dada por:

$$\lambda(h) = \frac{1}{2N(h)} \sum [y_s \cdot y_{(s+h)}]^2$$

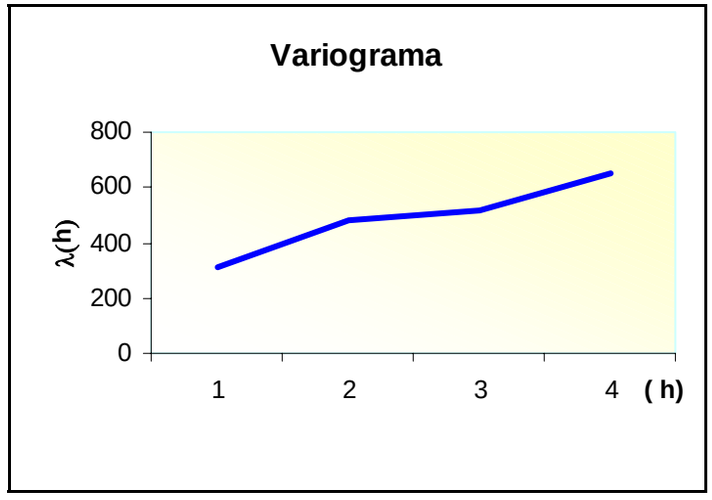
A estatística λ(h) equivale à metade da variância de diferenças entre observações separadas por uma distância h. O fator $\frac{1}{2}$ refere-se ao fato do interesse residir sobre a distância perpendicular dos pontos em relação à reta de 45^o. O conceito de semivariância é próprio da estatística espacial e, portanto, novo em relação à estatística clássica. Refere-se a uma medida natural da dispersão da nuvem de pontos. O prefixo semi advém do fator ($\frac{1}{2}$) da fórmula. Diferentemente da autocorrelação e da autocovariância, a semivariância cresce com o aumento da distância entre as observações, evidenciando que a nuvem de pontos torna-se mais dispersa ou difusa. Valores baixos de λ(h) indicam menor variabilidade, ou seja, maior similaridade.

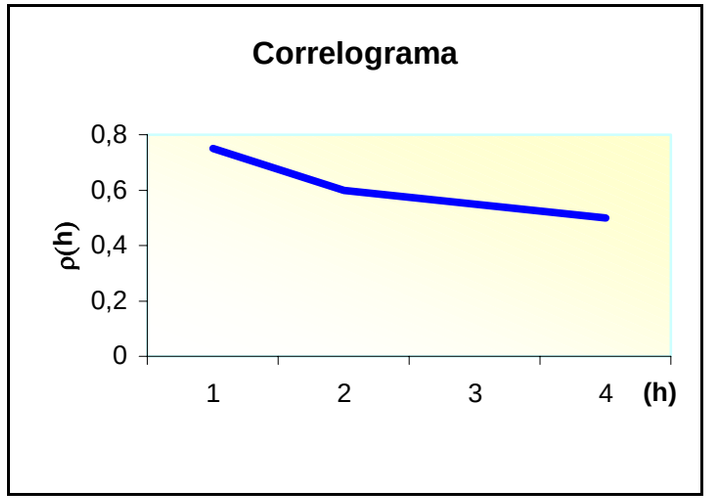
A relação entre a semivariância λ(h) e a distância h é denominada semivariograma ou simplesmente variograma. O termo semivariograma é mais correto teoricamente, entretanto, o jargão variograma tem uso mais amplo.

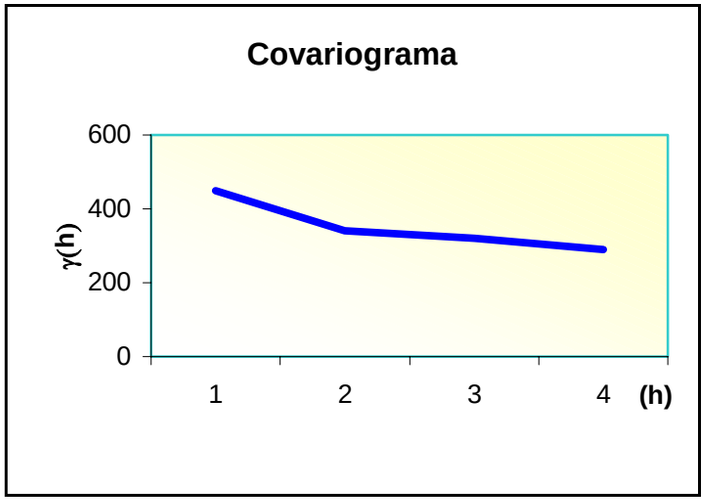
Variogramas, correlogramas e covariograma para a descrição da continuidade espacial podem ser obtidas a partir da semivariância, autocorrelação e autocovariância, respectivamente, associados a diferentes distâncias h em uma determinada direção, por exemplo, conforme a Tabela a seguir, segundo a direção norte-sul:

h	λ(h)	ρ(h)	γ(h)
(0,1)	310	0,75	450
(0,2)	480	0,60	340
(0,3)	520	0,55	320
(0,4)	650	0,50	290

O variograma, o correlograma e a função de autocovariância associados são:







Verifica-se pelo correlograma e pelo covariograma uma redução da dependência espacial com o aumento da distância entre as observações. Verifica-se pelo variograma um aumento da dispersão dos dados com o aumento da distância entre as observações.

Na prática, correlogramas e variogramas devem considerar mais de uma direção e considerar pelo menos os sentidos das linhas e colunas, verificando como os dados se correlacionam nestas duas direções.

Os três métodos são adequados para descrição da continuidade espacial. Entretanto, para propósitos de estimação, a função de covariância é mais relevante, por exemplo, é essencial na estruturação da matriz de covariância dos erros, quando os erros são correlacionados. A função de covariância e o correlograma são mais resistentes a valores erráticos do que o variograma, pois levam em conta as médias e variâncias dos “lags”, respectivamente. Em análise de séries temporais, as funções de covariância tendem a ser preferidas, uma vez que as mesmas têm a característica de diminuir com o aumento da distância. Para valores de h = 0, a autocovariância equivale simplesmente a variância σ^2 da amostra e a autocorrelação equivale a 1.

1.3 Isotropia, Alcance e Efeito Pepita na Análise Variográfica

A semivariância $\lambda(h)$ pode ser avaliada somente à distância h correspondente a múltiplos do espaçamento entre pontos de amostragem ao longo da direção considerada. O vetor h, apresentando-se infinitamente pequeno, faz com que a variância e a covariância ou autocovariância se tornem muito próximas. Para valores grandes de h, a covariância diminuirá ao passo que a variância aumentará. Dessa forma, a semivariância distribui-se de 0, quando h = 0, até um valor igual à variância das observações, para um valor alto de h. A distância na qual $\lambda(h)$ atinge um patamar igual à variância dos dados, patamar este denominado **soleira** (sill), é chamada **alcance**. A soleira é simbolizada por *C* e o alcance por α.

A variável regionalizada é composta de duas partes: a tendência e o resíduo. A tendência é o valor esperado da variável regionalizada em um determinado ponto x_i, que equivale à média ponderada de todos os pontos em torno de uma vizinhança x_i. Subtraindo a tendência, da variável regionalizada, os próprios resíduos serão a variável regionalizada estacionária. A construção do semivariograma pode ser baseada nos dados reais ou nos resíduos e faz parte da análise estrutural em geoestatística.

Os semivariogramas expressam o comportamento espacial da variável regionalizada e informam sobre:

(i) **padrão de variabilidade nas várias direções**: tem-se uma isotropia quando o padrão de variabilidade é o mesmo em todas as direções, gerando semivariograma omnidirecional; tem-se anisotropia quando o padrão de variabilidade difere em função das direções, requerendo semivariogramas direcionais;

(ii) **efeito pepita ou nugget**: ocorre quando para h = 0, a semivariância $\lambda(h)$ já apresenta algum valor, quando deveria ser nula revelando similaridade absoluta à distância nula.

O efeito pepita é simbolizado por C_o e pode ser atribuído a erros de medição ou ao fato dos dados não terem sido coletados a intervalos suficientemente pequenos para exibir o comportamento espacial do fenômeno estudado. O efeito pepita significativo denota que há uma grande variabilidade ou dissimilaridade à pequena escala ou distâncias extremamente pequenas. Tal efeito causa uma descontinuidade na origem do variograma e um salto vertical do valor 0 na origem para o valor da semivariância no variograma.

O efeito C_o mede fundamentalmente duas parcelas da variabilidade total do fenômeno: (a) a variabilidade correspondente a uma pequena escala não abrangida pela malha de amostragem; (b) a variabilidade induzida por erros não sistemáticos de amostragem, os quais acrescentam um ruído branco ou aleatório.

(iii) **forma da variabilidade espacial**: em manchas, em gradientes ou completamente aleatória.

O gráfico formado por $[(h), \lambda(h)]$ é denominado semivariograma experimental. O grau de aleatoriedade presente nos dados é dado pela expressão $r = C_o/C$ e pode ser interpretado da seguinte maneira:

r < 0,15: componente aleatória pequena;

0,15 < r < 0,30: componente aleatória significativa;

r > 0,30: componente aleatória muito significativa.

Após a obtenção do semivariograma experimental, é necessário ajustá-lo a um modelo teórico. Ajustar um semivariograma através de uma curva média permite inferir sobre o comportamento de $\lambda(h)$ representativo para toda a área e gama de valores de h.

Para a descrição da continuidade espacial, a construção do variograma é suficiente. Entretanto, para propósitos de estimação e predição, o ajuste de modelos ao variograma é necessário.

Para malhas regulares, como a dos experimentos de campo, o semivariograma experimental é obtido conforme os seguintes passos (Ribeiro Júnior, 1995; Duarte, 2000):

(i) fixa-se uma distância h ou “lag”;

(ii) formam-se todos os pontos separados pela distância h;

(iii) aplica-se a expressão do estimador para se obter a semivariância associada à distância h;

(iv) toma-se outra distância ou “lag” e repetem-se os passos de (i) a (iii), o que deve ser feito até uma distância máxima de interesse;

(v) obtém-se o semivariograma, plotando-se os pontos formados pelas distâncias no eixo x e pelas semivariâncias estimadas, no eixo y.

Dentre os modelos teóricos de semivariogramas, os principais são o exponencial, o esférico e o normal.

No ajustamento de variogramas, pode-se iniciar com o omnidirecional, no qual todas as direções possíveis são combinadas em um único variograma. Tal pode ser entendido como uma média aproximada dos vários variogramas direcionais. A obtenção do variograma omnidirecional não implica assumir definitivamente que a continuidade espacial é a mesma em todas as direções. Ele pode ser usado apenas como ponto de partida na obtenção de parâmetros necessários ao ajuste de variogramas, tais como a descoberta dos parâmetros de distância (h) que propiciam uma estrutura mais nítida.

O variograma omnidirecional é baseado em maior número de pares de pontos que os variogramas direcionais. Assim, provavelmente propicie uma estrutura mais nítida. Se tal variograma não produz uma estrutura espacial clara, os direcionais provavelmente também não apresentarão. Após a obtenção de um variograma omnidirecional claro, pode-se explorar o padrão de anisotropia e obter os vários variogramas direcionais.

1.4 Variogramas em Experimentos de Campo

A variabilidade espacial pode ser caracterizada como:

(i) **Contínua**: refletindo padrões similares devido a efeitos de solo e clima. Neste caso, a variação contínua espacial pode aparecer sob a forma de: (a) tendência local ou manchas; (b) tendência global ou gradientes sobre todo o local.

(ii) **Descontínua**: como reflexo de práticas culturais ou efeitos de mensuração.

(iii) **Aleatória**: devido à heterogeneidade de microambiente.

A variabilidade espacial pode ser estudada basicamente por meio de duas classes de métodos: os métodos de análise de séries temporais e os métodos geoestatísticos. Por meio dos métodos de análise de séries temporais, Gleeson e Cullis (1987) sugeriram o uso de um modelo

auto-regressivo de primeira ordem (AR1) para modelar os resíduos em uma dimensão do espaço e o uso do método REML para estimar os parâmetros do modelo. Em um modelo AR1, a autocorrelação [$\rho(Y_i , Y_j)$] entre as observações Y_i e Y_j é uma função potência da distância entre as observações, de forma que $\rho(Y_i , Y_j) = \rho^{|i-j|}$, em que i e j referem-se às coordenadas espaciais e ρ é o coeficiente de autocorrelação. Um modelo auto-regressivo de primeira ordem indica que somente a correlação entre observações imediatamente vizinhas são diretamente especificadas. Correlações entre vizinhos mais distantes surgem somente como consequências dessas correlações de primeira ordem. Modelos de ordem mais elevada (por exemplo um AR2) podem ser especificados, nos quais observações não adjascentes podem apresentar dependência direta, além daquela indireta contemplada pelo modelo AR1.

Cullis e Gleeson (1991) estenderam o modelo para considerar a variabilidade em duas dimensões do espaço considerando processos (AR1⊗AR1) separáveis em duas direções: linhas e colunas. Neste modelo, a autocorrelação é dada por: $\rho\left(Y_{i,j} , Y_{k,\ell}\right) = \rho_{lin}^{[i-k]} \ \rho_{col}^{[j-\ell]}$ para observações com coordenadas $i,\ j\ e\ k,\ \ell$ referentes a linhas e colunas, respectivamente.

Zimmerman e Harville (1991) demonstraram uma relação muito próxima entre as várias abordagens para o estudo da variabilidade espacial e propuseram uma metodologia na qual a tendência global é modelada usando efeitos fixos e a tendência local modelada por meio de uma estrutura de correlação baseada em geoestatística. Isto é coerente pois manchas não tem como ser consideradas adequadamente usando efeitos fixos (blocos, por exemplo), ao contrário dos gradientes.

Gilmour et al. (1977) e Cullis et al. (1998) sugerem um procedimento de ajuste de modelos em que o variograma amostral, obtido a partir de uma análise residual inicial com base no modelo AR1⊗AR1, é usado para detectar tendência global e efeitos extras (variação descontínua) tais como de linhas e colunas. Se estes efeitos extras forem identificados, o modelo pode ser estendido pela adição destes efeitos. Nesta abordagem, a tendência local é modelada pelo próprio processo AR1⊗AR1, que tem se mostrado adequado para tal fim. Os efeitos extras e de tendência global têm sido ajustados como efeitos fixos (Smith, Cullis e Thompson, 2001).

Assim, os variogramas são essenciais no ajuste de modelos e verificação da eficácia dos modelos estendidos. Idealmente, devem-se ter variogramas estacionários, ou seja, aqueles que atingem um patamar ou soleira (*sill*), para que a modelagem dos erros esteja adequada (neste caso, os variogramas têm uma relação simples e direta com a estrutura de autocorrelação). Séries não estacionárias tornam-se estacionárias removendo-se algumas tendências nos dados. Uma estacionariedade de segunda ordem é suficiente, condição que é normalmente atendida desde que o variograma seja típico (com uma parte crescente e um patamar) e estacionário – hipótese intrínseca (Duarte, 2000).

Com $\rho_{lin}\ \rho_{col} = 0$, o variograma é liso e não apresenta nenhuma estrutura espacial ou padrão distinguível. Com o aumento de $\rho_{lin}\ e\ \rho_{col}$ aparecem as manchas a partir do valor 0,3 para a autocorrelação. Nos variogramas, a tendência local é detectada pela não estacionariedade e as variações extras pelas ondulações ou rugas. A tendência global pode ser modelada pelo ajuste de polinômios quadráticos nas coordenadas espaciais, ou, slipnes cúbicas alisadas, quando não há interação entre termos de linhas e colunas.

O uso de variogramas e sua análise por vários modelos espaciais auto-regressivos com ou sem extensão permite a escolha de modelos adequados, assim como se faz com o uso dos métodos geoestatísticos.

2 MÉTODOS DE ANÁLISE ESPACIAL DE EXPERIMENTOS

A análise tradicional de experimentos de campo assume que todas as observações tomadas em posições adjacentes são não correlacionadas. Assim, a matriz de covariância residual é modelada como uma matriz diagonal, ou seja, com os erros assumidos como independentes. Também, a posição dos tratamentos no campo, ou seja, a distribuição espacial dos mesmos é ignorada. Entretanto, a dependência espacial pode existir e contribuir para o aumento da variação residual, de forma que pode ser importante considerá-la nas análises.

A casualização concorre para a neutralização dos efeitos da correlação espacial e, portanto, para a geração de uma análise de variância fidedigna. Entretanto, embora a teoria da casualização enfatize a neutralização da correlação espacial, tal neutralização é mais eficiente quando se usam modelos espaciais. Também as formas de controle local baseadas em blocagem podem ser ineficientes para tratar de problemas de gradientes ambientais e mesmo os blocos incompletos podem não permitir uma avaliação completa dos efeitos espaciais. Além disso, a blocagem é realizada antes da implantação dos experimentos, de forma que percebe-se muitas vezes, por ocasião da coleta dos dados experimentais, a presença de manchas ou gradientes ambientais dentro dos experimentos, os quais não foram considerados adequadamente pelos blocos delineados *a priori*. Nesta situação, somente as técnicas de análise espacial permitem contornar a questão e propiciar uma seleção acurada, através de blocagem *a posteriori* ou através da flexibilização da matriz R baseados nos próprios dados experimentais, conforme realizado por Duarte (2000). A variabilidade ou heterogeneidade espacial associada à fertilidade e estrutura do solo, umidade, interceptação de luz e outros fatores ambientais contribuem para o aumento da variação residual. Assim, é importante controlar, por delineamento ou por análise, a variação residual espacial ou tendência em fertilidade.

Além dos delineamentos experimentais, outras formas de controle local e aumento da precisão experimental referem-se aos procedimentos de análise espacial, os quais podem ser agrupados em duas classes principais: ***métodos geoestatísticos*** (Cressie, 1993; Grondona e Cressie, 1991; Zimmerman e Harville, 1991) e ***métodos de análise de séries temporais*** (Gleeson e Cullis, 1987; Martin, 1990; Cullis e Gleeson, 1991; Gilmour, Cullis e Verbyla, 1997; Gilmour et al., 1998; Cullis et al. 1998; Smith, Cullis e Thompson, 2001) usando estimativas REML de componentes de variância (Cooper e Thompson, 1977). Estes últimos consideram os erros por meio de um processo auto-regressivo integrado de médias móveis (ARIMA (p, q, d)) que pode ser aplicado a duas dimensões: linhas e colunas. Tal modelo estendido é da forma ARIMA (p_1, d_1, q_1) x ARIMA (p_2, d_2, q_2) (Cullis e Gleeson, 1991; Martin, 1990). Estes modelos são denominados modelos com erros nas variáveis e consideram um efeito de tendência (ξ) mais um erro η independente ou efeito pepita. Assim, o vetor de erros é particionado em e = ξ + η. Os modelos de análise tradicionais não incluem o componente ξ.

A variância dos resíduos é dada por Var(e) = Var (ξ + η)= R = Σ = $\sigma_{\xi}^{2}[\sum_{c}(\Phi_{c})\otimes\sum_{r}(\Phi_{r})]+I\sigma_{\eta}^{2}$, em que σ_{ξ}^{2} é a variância devida a tendência e σ_{η}^{2} é a variância dos resíduos não correlacionados. As matrizes $\sum_{c}(\Phi_{c})$ e $\sum_{r}(\Phi_{r})$ referem-se a matrizes de correlação auto-regressivas de primeira ordem com parâmetros de autocorrelação Φ_{c} e Φ_{r} e ordem igual ao número de colunas e número de linhas, respectivamente. Assim, ξ é modelado como um processo auto-regressivo separável de primeira ordem (AR1 x AR1) com matriz de covariância $Var(\xi)=\sigma_{\xi}^{2}[\sum_{c}(\Phi_{c})\otimes\sum_{r}(\Phi_{r})]=H\sigma_{\xi}^{2}$, em que $H=[\sum_{c}(\Phi_{c})\otimes\sum_{r}(\Phi_{r})]$. Em culturas anuais e em ausência de conhecimento da correta estrutura de correlação, Gilmour, Cullis e Verbyla (1997) sugerem a modelagem de ξ como um processo (AR1 x AR1). Este processo autoregressivo em duas dimensões tem mostrado eficiência em várias situações (Grondona et al., 1996; Gilmour, Cullis e Verbyla, 1997; Cullis et al., 1998; Qiao et al., 2000; Costa e Silva et al., 2001; Resende e Sturion, 2001; 2002; Smith, Cullis e Thompson, 2001; Stringer e Cullis, 2002; Resende e Thompson, 2003). Os métodos ARIMA de Gleeson e Cullis (1987), Martin (1990) e Cullis e Gleeson (1991) abrangem os métodos de vizinhança (NN) de Papadakis (1937), iterativo de vizinhança de Papadakis (Papadakis, 1970; Bartlett, 1978) e outros métodos prévios (Papadakis, 1984; Bartelett, 1938; Atkinson, 1969; Wilkinson et al., 1983; Green et al., 1985; Besag e Kempton, 1986; Williams, 1986) de análise de vizinhança.

A geoestatística consiste basicamente de variografia e krigagem. A variografia usa variogramas para caracterizar e modelar a variação espacial. A krigagem usa a variação modelada para predizer valores, tais quais os BLUPs de erros correlacionados. O variograma usa semivariâncias e pode ser usado em ambos os métodos de análise espacial: geoestatística e modelos de séries temporais. Pela geoestatística, o modelo padrão para ajuste de uma função ao variograma experimental em ensaios de campo é o exponencial. Cressie (1993) tentou ajustar várias classes de modelos ao variograma experimental em vários ensaios de campo e concluiu que nenhum produziu melhor ajuste do que o modelo exponencial.

Os procedimentos geoestatísticos consideram a heterogeneidade espacial de forma direta por meio da inclusão dos efeitos de tendência e correlação residual na modelagem da matriz de covariância residual. De acordo com Gleeson (1997), a abordagem geoestatística de Zimmerman e Harville (1991) denominada

modelo linear de campo aleatório é equivalente ao ajuste de processos ARIMA separáveis e, de acordo com Gilmour, Thompson e Cullis (1995), é equivalente a um processo (AR1 x AR1) com erro independente η.

Como o modelo associado ao variograma é exponencial, os resíduos podem ser interpretados como uma realização de um processo auto-regressivo de primeira ordem (AR1). Isto faz sentido uma vez que o modelo AR1 projeta a auto-correlação para lags distantes, como uma função potência da distância entre plantas. O modelo exponencial faz o mesmo. Entretanto, os modelos geoestatísticos muitas vezes assumem isotropia, o que pode ser inadequado para modelar a estrutura de variâncias nos experimentos de campo. Considerando uma estrutura de covariância exponencial direcional (anisotrópica) no modelo geoestatístico, foi demonstrada (Gilmour, Cullis e Verbyla, 1997) formalmente a equivalência entre a modelagem geoestatística exponencial e o modelo separável AR1 x AR1 para experimentos de campo. Em função desta equivalência e da facilidade em ajustar modelos anisotrópicos pela modelagem ARIMA, esta tem sido preferida. Adicionalmente, a separabilidade resulta em maior eficiência computacional em termos de tempo. Dessa forma, os modelos ARIMA devem ser preferidos já que englobam também as demais metodologias de análise de vizinhança apresentadas na literatura, tais quais os métodos de Papadakis e modificações (Gilmour, Thompson e Cullis, 1995). Adicionalmente, baseado em resultados recentes, Cullis (2005) relata *"the standard practice of plotting empirical variograms e selecting the model can lead to severe model misspecification"*. Tal autor relata que a dependência espacial muda com a direção de uma maneira elíptica: anisotropia geométrica. Sugere então o uso de funções de covariância de Matern como melhor abordagem.

3 MODELO LINEAR MISTO ESPACIAL COM ERROS AR1 X AR1

Considerando um modelo com erros espacialmente correlacionados para a análise de um teste de progênies em espécie perene, tem-se $Y_{ijk} = \mu + p_i + b_j + c_{ij} + \xi_{ijk} + \eta_{ijk}$, em que *p*, *b*, *c*, ξ e η são os efeitos de progênie, bloco (fixo), parcela e erros correlacionados e independentes dentro de parcela, respectivamente. O índice multi-efeitos espacial (IMEE) para a predição dos efeitos genéticos aditivos é dado por $\hat{a}_{ijK} = b_1 p_i + b_2 c_{ij} + b_3 \xi_{ijk} + b_4 \eta_{ijk}$ (Resende e Thompson, 2003). Na situação em que o componente ξ_{ijk} é significativo (erro correlacionado de alta magnitude), este índice expandido é superior ao tradicional índice multi-efeitos (IME) apresentado por Resende e Higa (1994). O modelo acima, em versão matricial (modelo linear misto) é dado por

$$y = Xb + Za + Wc + \varepsilon$$
$$= Xb + Za + Wc + \xi + \eta$$, em que *a* e *c* são os efeitos aleatórios genéticos (aditivos individuais) e de parcelas, respectivamente.

O IMEE em sua versão BLUP é dado por

$$\begin{bmatrix} X'X & X'Z & X'W & X'I \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'W & Z'I \\ W'X & W'Z^* & W'W + I\lambda_2 & W'I \\ I'X & I'Z^* & I'W & I'I + H^{-1}\lambda_3 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{a} \\ \tilde{c} \\ \tilde{\xi} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ I'y \end{bmatrix}$$, em que:

$$\lambda_1 = \frac{\sigma_\eta^2}{\sigma_a^2}; \qquad \lambda_2 = \frac{\sigma_\eta^2}{\sigma_c^2} \qquad \lambda_3 = \frac{\sigma_\eta^2}{\sigma_\xi^2}.$$

A e H são as matrizes de correlação para os efeitos *a* e ξ, respectivamente.

A inversão de H é dada por $H^{-1} = [\sum_c^{-1}(\Phi_c) \otimes \sum_r^{-1}(\Phi_r)]$. A estimação da variância do erro correlacionado via REML pode ser dada por $\hat{\sigma}_\xi^2 = [\tilde{\xi}' H^{-1} \tilde{\xi} + \hat{\sigma}_\eta^2 \; tr \; (H^{-1} C^{44})] / N$, em que C^{44} advém da inversa da matriz dos coeficientes e N é o número total de dados.

O componente dentro de progênies de meios irmãos para o IMEE equivale a $\hat{h}_{adj}^2 (y - X\hat{b} - Z_1 \hat{a}_g - W\hat{c} - \hat{\xi})$, enquanto que para o IME tradicional equivale a $\hat{h}_d^2 (y - X\hat{b} - Z_1 \hat{a}_g - W\hat{c})$. Verifica-se que $\hat{h}_{adj}^2$ é uma herdabilidade ajustada dentro de progênies ($\hat{h}_{dadj}^2 = (0.75\hat{\sigma}_a^2)/(0.75\hat{\sigma}_a^2 + \hat{\sigma}_\eta^2)$) e que apresenta magnitude igual ou superior a $\hat{h}_d^2$ do IME tradicional ($\hat{h}_d^2 = (0.75\hat{\sigma}_a^2)/(0.75\hat{\sigma}_a^2 + \sigma_\xi^2 + \hat{\sigma}_\eta^2)$).

Para a verificação da superioridade da análise espacial, algumas quantidades podem ser empregadas. Em geral, o modelo de melhor ajuste é aquele com menor erro, ou seja, aquele com menor variância do erro. O modelo com menor variância do erro é o modelo com maior determinação e maior Log L (logaritmo do máximo da função de verossimilhança restrita). Entretanto, é preciso verificar se a diminuição da variância do erro às custas de um maior número de parâmetros, é significativa. Isto pode ser feito via REMLRT (teste da razão de verossimilhança)

usando os Log L dos modelos em comparação. No entanto, tal teste não é capaz de indicar, em termos genéticos, a superioridade ou eficiência de um modelo. As seguintes estatísticas podem ser usadas como medidas de eficiência: (i) relação $\sigma_\eta^2 / \sigma_\varepsilon^2$ referente aos modelos espacial e não espacial; (ii) função associada ao fator de *shrinkage* para os efeitos genéticos, $\lambda_{1e} = \sigma_\eta^2 / \sigma_a^2$ e $\lambda_{1n} = \sigma_\varepsilon^2 / \sigma_a^2$, para os modelos espaciais e não espaciais, respectivamente; (iii) herdabilidade ajustada para todos os efeitos do modelo, $\hat{h}_{adj}^2 = (\hat{\sigma}_a^2)/(\hat{\sigma}_a^2 + \hat{\sigma}_\eta^2) = 1/(1+\lambda_1)$.

Assim, quanto maior a herdabilidade ajustada, menor o *shrinkage* (função de $\lambda_1 = (1 - \hat{h}_{adj}^2)/\hat{h}_{adj}^2$), maior a acurácia seletiva e maior o ganho genético. Também, quanto menor a variância do erro, maior a herdabilidade ajustada desde que a variância genética estimada não decresça com o modelo espacial. Assim, a eficiência pode ser computada como uma razão entre as herdabilidades ajustadas dos dois modelos (Resende, Thompson e Welham, 2003). Estimativas de variância genética menores pelo modelo espacial revelam inadequação do mesmo e isto geralmente ocorre quando as estimativas dos coeficientes de auto-correlação não diferem estatisticamente de 1. Neste caso, a tendência pode ser efetivamente removida pelo delineamento experimental, por meio dos efeitos de blocos e parcelas.

O índice multi-efeitos espacial pode ser estendido pela incorporação dos efeitos de competição, gerando um IMEEC. Neste caso, basta desdobrar o efeito de tratamento em que $a = Z\tau + NZ\phi$, em que τ é o efeito genético direto no próprio tratamento e ϕ é o efeito genético indireto no vizinho (ver Capítulo 6). A análise espacial pode ser útil também no aumento da eficiência da seleção massal em plantios sem delineamento experimental. Neste caso, um modelo com efeitos de linhas e colunas mais os erros correlacionado e independente pode ser ajustado. O erro independente será usado então para a seleção, em lugar do valor fenotípico bruto. Isto é vantajoso pois $\hat{\eta}$ conterá os efeitos genético e ambiental não correlacionados, após correção para os efeitos de linhas, colunas e $\hat{\xi}$.

Um outro procedimento melhorado em relação ao índice multi-efeitos tradicional de Resende e Higa (1994) é o índice multi-efeitos, considerando um delineamento em linha e coluna (IMELC). O IMELC considera simultaneamente dois sistemas de blocagem, isto é, no sentido das linhas e colunas, visando considerar gradientes de fertilidade em duas direções, assim como o faz o

delineamento em quadrado latino. No caso, o modelo linear associado é dado por $Y_{(i)jkl} = \mu + p_{(i)} + l_j + \kappa_k + c_{jk} + \delta_{(i)jkl}$, em que *p*, *l*, κ, *c* e δ são os efeitos de progênie, linha, coluna, parcela e erro dentro de parcela, respectivamente. Conforme Resende (2004), o IMELC para a predição dos efeitos genéticos aditivos é dado por $\hat{a}_{(i)jKl} = b_1 p_{(i)} + b_2 l_j + b_3 \kappa_k + b_4 c_{jk} + b_5 \delta_{(i)jkl}$.

$$\hat{a}_{(i)jKl} = b_1\left(\overline{Y}_{(i)} - \overline{Y}_{...}\right) + b_2\left(\overline{Y}_{j..} - \overline{Y}_{...}\right) + b_3\left(\overline{Y}_{.k.} - \overline{Y}_{...}\right) + b_4\left(\overline{Y}_{jk.} - \overline{Y}_{(i)} - \overline{Y}_{j..} - \overline{Y}_{.k.} + 2\overline{Y}_{...}\right) + b_5\left(Y_{(i)jkl} - \overline{Y}_{jk.}\right)$$

No entanto, a utilização do delineamento em linha e coluna (DLC) em experimentos instalados em blocos ao acaso (DBC) caracteriza um delineamento não ortogonal e o IMELC deve ser estabelecido via BLUP, conforme incorporado no *software* Selegen-REML/BLUP. O IMELC foi aplicado eficientemente (superioridade cerca de 5 % em termos da herdabilidade ajustada) em DBC com várias plantas por parcela, onde as plantas dentro de parcela estavam dispostas de maneira perpendicular ao sentido dos blocos (Resende, Thompson e Welham, 2003). No caso, cada bloco continha seis linhas (devido a seis plantas por parcela). Com o DLC, o tamanho de cada bloco passou a equivaler a uma linha e aumentou-se o número de blocos (então representados pelas linhas). É importante relatar que, nesse caso, cada coluna é incompleta e portanto deve ser ajustada como efeito aleatório. O efeito de linha, no caso, pode ser ajustado como fixo ou aleatório. O IMELC é um melhoramento do IME sem usar a análise espacial, sendo intermediário entre o IME e o IMEE.

4 ESTRUTURA ESPACIAL MULTIVARIADA E COM MEDIDAS REPETIDAS

A seguir será detalhada um pouco mais a estrutura autoregressiva AR1 x AR1. Considerando um experimento com forma retangular em uma malha de c colunas e r linhas, os resíduos podem ser arranjados em uma matriz de forma que podem ser considerados como correlacionados dentro de colunas e de linhas. Escrevendo esses resíduos em um vetor, seguindo a ordem de campo (colocando cada coluna da grade experimental uma sob a outra), a variância dos resíduos é dada por $Var(\varepsilon) = Var(\xi + \eta) = R = \Sigma = \sigma^2_\xi[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)] + I\sigma^2_\eta$, em que σ^2_ξ é a variância devida à tendência local e σ^2_η é a variância dos resíduos independentes.

As matrizes $\sum_c(\Phi_c)$ *and* $\sum_r(\Phi_r)$ referem-se a matrizes autoregressivas de primeira ordem com parâmetros de autocorrelação Φ_c e Φ_r e ordem igual ao número de colunas e de linhas, respectivamente.

Uma matriz de autocorrelação de primeira ordem AR1(ρ) é da forma (para quatro colunas ou linhas):

$$\sum(\rho)=\begin{bmatrix} 1 & \rho^{|t_2-t_1|} & \rho^{|t_3-t1|} & \rho^{|t_4-t_1|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_3-t_2|} & \rho^{|t_4-t_2|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 & \rho^{|t_4-t_3|} \\ \rho^{|t_4-t_1|} & \rho^{|t_4-t_2|} & a^{|t_4-t_3|} & 1 \end{bmatrix}=\begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

Outra formulação que pode ser usada é R = S + R*, em que $S=Var(\xi)=\sigma^2_{\xi}[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$ e $R^*=I\sigma^2_{\eta}$. Para o caso de várias variáveis ou medidas repetidas tomadas em cada unidade experimental, a estrutura espacial multivariada é dada por (Resende e Thompson, 2003; Resende, Thompson e Welham, 2006) considerando o caso bivariado:

$S^{-1}=S_o^{-1}\otimes H^{-1}=\begin{bmatrix} H\sigma^2_{\xi_1} & H\sigma_{\xi_{12}} \\ H\sigma_{\xi_{12}} & H\sigma^2_{\xi_2} \end{bmatrix}^{-1}$, em que:

$$S_O=\begin{bmatrix} \sigma^2_{\xi_1} & \sigma_{\xi_{12}} \\ \sigma_{\xi_{12}} & \sigma^2_{\xi_2} \end{bmatrix} \qquad H=[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$$

No caso de medidas repetidas, os demais fatores (tratamentos, parcela, etc) do modelo podem demandar também uma modelagem multivariada. Por exemplo, os efeitos de tratamentos podem ser ajustados por modelos autoregressivos (ARH) e de ante-dependência (SAD) com variâncias heterogêneas, e essa modelagem deve ser realizada simultaneamente à análise espacial. Quando os efeitos espaciais são importantes, a combinação de modelos autoregressivos para os resíduos e modelos SAD ou ARH para os efeitos de tratamentos podem ser eficientes, conforme verificado por Resende e Thompson (2003) e Resende, Thompson e Welham (2006). A análise espacial associada à análise conjunta de vários experimentos é discutida em detalhes por Resende e Thompson (2004).

5 RESULTADOS PRÁTICOS DA ANÁLISE ESPACIAL

Resultados referentes à análise espacial de cinco experimentos são apresentados (Tabela 13) e discutidos visando elucidar as interpretações da análise espacial.

Tabela 13. Máximo do logaritmo da função de verossimilhança residual (Log L) e estimativas da variância genética entre tratamentos ($\hat{\sigma}^2_\tau$), variância residual ($\hat{\sigma}^2$), herdabilidade ajustada ($\hat{h}^2_{adj}$), proporção de variâncias residuais associadas aos modelos espacial e não espacial ($\hat{\sigma}^2_s / \hat{\sigma}^2_{ns}$) e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr). Dados associados a cinco diferentes experimentos.

Conjunto de Dados – Modelo	**Log L**	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2$	$\hat{h}^2_{adj}$	$\hat{\sigma}^2_s / \hat{\sigma}^2_{ns}$	**ARc**	**ARr**
Experimento 1							
E. grandis - Tradicional	-19487.4	15.414	24.827	0.383	-	-	-
E. grandis - Espacial	-19407.2	15.514	24.607	0.387	0.991	0.81*	0.99*
Experimento 2							
Erva-mate-Tradicional	1552.89	0.0110	0.2214	0.1905	-	-	-
Erva-mate -Espacial	2127.24	0.0134	0.1492	0.3296	0.674	0.70*	0.75*
Experimento 3							
Cana de Açúcar - Tradicional	-1023.77	685.98	228.73	0.7499	-	-	-
Cana de Açúcar – Espacial	-1023.05	684.17	229.26	0.7490	1.000	-0.12^{ns}	0.035^{ns}
Experimento 4	-6584.67	1.0403	18.255	0.2156	-	-	-
Pínus -Tradicional	-6559.19	0.9621	17.891	0.2040	0.980	-0.10*	-0.13*
Pínus - Espacial							
Experimento 5	-2940.18	37.25	601.44	0.2333	-	-	-
E. maculata -Tradicional	-2935.12	35.80	596.61	0.2264	0.991	0.10*	0.10*
E. maculata - Espacial							

O primeiro experimento refere-se a dados de crescimento em eucalipto aos dois anos de idade. O delineamento empregado foi o de blocos incompletos (látice). Os resultados foram: (a) herdabilidade ajustada da análise tradicional igual a 0,38; (b) herdabilidade ajustada da análise espacial igual a 0,39; (c) coeficientes de autocorrelação residual nas linhas e colunas iguais a 0,81 e 0,99, respectivamente; (d) relação entre variâncias residuais (erros não correlacionados) pelas duas análises igual a 0,99; (e) variâncias entre blocos (Vb) do látice pelas análises tradicional e espacial iguais a 1,58 e 0,02, respectivamente (Tabela 14). As seguintes conclusões podem ser emitidas: (i) não existe efeitos de competição já que os coeficientes de autocorrelação foram positivos e de alta magnitude (tal resultado já era esperado dado a idade das plantas avaliadas); (ii) não houve eficiência ou melhoramento da análise com o uso da abordagem espacial já que as herdabilidades ajustadas obtidas pelas duas abordagens foram praticamente iguais e a relação entre variâncias residuais tendeu a 1; (iii) esta não eficiência pode ser explicada pelos coeficientes de autocorrelação tenderem a 1, evidenciando tendência linear a qual pode ser efetivamente removida pelo delineamento experimental por meio dos efeitos de blocos; (iv) a afirmativa anterior é corroborada pelos valores elevados de Vb apresentados na análise tradicional (removendo a tendência) e valor nulo na análise espacial, ou seja, a blocagem e a análise espacial desempenharam o mesmo papel.

Tabela 14. Máximo do logaritmo da função de verossimilhança residual (Log L) e estimativas da variância genética entre tratamentos ($\hat{\sigma}^2_\tau$), variância residual ($\hat{\sigma}^2$), variância dos erros correlacionados ($\hat{\sigma}^2_\xi$), variância entre blocos ($\hat{\sigma}^2_b$), proporção de variâncias residuais associadas aos modelos espacial e não espacial ($\hat{\sigma}^2_s / \hat{\sigma}^2_{ns}$) e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr). Dados associados ao experimento 1 de *E. grandis.*

Modelo de Análise	**Log L**	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2$	$\hat{\sigma}^2_\xi$	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_s / \hat{\sigma}^2_{ns}$	**ARc**	**Arr**
E. grandis -Traditional	-19487.4	15.414	24.827	-	1.576	1.000	-	-
E. grandis - Espacial + η	-19407.2	15.514	24.607	6.686	0.023	0.991	0.81*	0.99*
E. grandis - Espacial	-19484.7	15.316	-	24.732	1.687	0.996	0.00^{ns}	-0.03^{ns}

O segundo experimento refere-se a dados de massa foliar em erva-mate. O delineamento empregado foi o de blocos completos. Os resultados foram: (a) herdabilidade ajustada da análise tradicional igual a 0,19; (b) herdabilidade ajustada da análise espacial igual a 0,33; (c) coeficientes de autocorrelação residual nas linhas e colunas iguais a 0,70 e 0,75, respectivamente; (d) relação entre variâncias residuais (erros não correlacionados) pelas análises espacial e tradicional igual a 0,67. As seguintes conclusões podem ser emitidas: (i) não existe efeitos de competição, já que os coeficientes de autocorrelação foram positivos e de alta magnitude (tal resultado já era esperado visto que as plantas são podadas anualmente e são bastante espaçadas); (ii) houve eficiência ou melhoramento da análise com o uso da abordagem espacial, já que a herdabilidade ajustada obtida pela análise espacial foi muito superior à obtida pela análise tradicional e a relação entre variâncias residuais distanciou de 1; (iii) esta eficiência pode ser explicada pelos coeficientes de autocorrelação intermediários entre 0 e 1, evidenciando uma tendência que não pode ser efetivamente removida pelo delineamento experimental por meio dos efeitos de blocos.

Os experimentos de 3 a 5 apresentam significativos efeitos de competição, conforme revelado pelos baixos e negativos coeficientes de auto-correlação residual e por isso serão melhor discutidos no Capítulo 6. O quinto experimento refere-se a dados de crescimento em eucalipto aos 14 anos de idade. O delineamento empregado foi o de blocos completos. Os resultados foram: (a) herdabilidade ajustada da análise tradicional igual a 0,23; (b) herdabilidade ajustada da análise espacial igual a 0,23; (c) coeficientes de autocorrelação residual nas linhas e colunas iguais a -0,10 e -0,10, respectivamente; (d) relação entre variâncias residuais (erros não correlacionados) pelas análises espacial e tradicional iguais a 0,99. As seguintes conclusões podem ser emitidas: (i) existem efeitos de competição já que os coeficientes de autocorrelação foram negativos (tal resultado já era esperado dado a elevada idade das plantas); (ii) não houve eficiência ou melhoramento da análise com o uso da abordagem espacial, já que as herdabilidades ajustadas obtidas pelas duas análises foram iguais e a relação entre variâncias residuais equivaleu a 0,99; (iii) esta não eficiência pode ser explicada pelos coeficientes de autocorrelação negativos e próximos a zero, evidenciando uma confundimento entre efeitos de competição e de tendência espacial. Com coeficientes de autocorrelação próximos a zero a análise espacial não propicia melhoria no ajuste de modelos. Os resultados completos das análises desses experimentos são apresentados por Resende e Thompson (2003).

Pelos resultados apresentados, constata-se que nem sempre a análise espacial apresenta maior eficiência do que a análise tradicional. Com o uso de um bom delineamento experimental (blocos incompletos quando o número de tratamentos é elevado) e pequeno tamanho de bloco (uma planta por parcela), a análise espacial provavelmente não será necessária.

Maiores detalhes teóricos e práticos da análise espacial de experimentos são apresentados por Duarte (2000), Resende e Sturion (2001), Resende (2002) e Resende e Thompson (2003) e não serão apresentados aqui.

6 ANÁLISE ESTATÍSTICA ESPACIAL DE QTL

Marcadores genéticos em ligação próxima com locos controladores de características quantitativas (QTL, que é um segmento cromossômico, não necessariamente apenas um gene) são usados para mapear QTLs e também na seleção auxiliada por marcadores em conjunto com informações fenotípicas advindas de experimentos de campo. As abordagens estatísticas para análise de QTL diferem em relação às suposições de efeitos fixos ou aleatórios de QTL. Alguns métodos assumem o QTL como efeito fixo e com número finito de alelos. Outros o assumem como efeito aleatório com um infinito número de alelos. Os métodos estatísticos que tratam o QTL como efeito fixo variam desde modelos simples de regressão a abordagens bayesianas. Tais modelos estatísticos são misturas de distribuições, em que o número de densidades componentes é determinado pelo número de genótipos do QTL. As suposições relativas ao número de alelos segregantes tem um grande efeito na formulação do modelo estatístico. Modelos de efeitos aleatórios oferecem uma abordagem menos parametrizada para o mapeamento. Em tal abordagem, os efeitos de QTL são assumidos como tendo distribuição normal.

Em um procedimento de mapeamento de QTL, inicialmente, análises de marcadores únicos são realizadas por meio de métodos estatísticos simples como a ANOVA, a ANOVA não paramétrica de Kruskal-Wallis, a estatística t de Student, a regressão linear simples, a máxima verossimilhança (LOD score). Estes procedimentos permitem a detecção de associação entre os marcadores e o caráter de interesse, sem usar informação de mapa genético. Isto é feito para cada marcador, contrastando as observações fenotípicas entre as classes de cada marcador. Tais classes são

tomadas como se fossem tratamentos a serem comparados. Posteriormente, o mapeamento por intervalo, considerando dois marcadores, pode ser feito visando à seleção de marcadores a serem usados como potenciais cofatores em uma análise de regressão múltipla do tipo *stepwise*. Também, o mapeamento por intervalo composto pode ser efetuado quando múltiplos QTLs estão ligados ao intervalo ou marcador considerados.

Em geral, os procedimentos de mapeamento têm usado diretamente os dados de campo para análise. Tais dados, em conjunto com a informação molecular são usados nos *softwares* padrões para mapeamento de QTL. Ou seja, não são rotineiramente usados valores genéticos preditos após a eliminação dos efeitos ambientais. Entretanto, é recomendável que o mapeamento seja baseado em valores genéticos preditos sob um modelo que contemple também os efeitos ambientais de escala global (locais, blocos), os efeitos ambientais de escala localizada (resíduo correlacionado ou espacial) e os efeitos de competição (se houverem). Também, em caso de experimentos envolvendo múltiplos locais, os efeitos da interação genótipo x ambiente devem também ser incluídos no modelo. No entanto, o procedimento ideal refere-se à inclusão simultânea dos efeitos dos marcadores no modelo de predição dos valores genéticos, de forma que o mapeamento seja realizado simultaneamente à predição. Este procedimento é superior devido ao fato de que os valores ou efeitos genéticos são preditos com diferentes precisões e também podem ser correlacionados devido à predição. Essas diferentes precisões e a correlação não são levadas em consideração quando não se adota a análise simultânea.

No contexto de QTLs como efeitos fixos, o método de regressão com dois marcadores flanqueadores permite naturalmente a análise combinada dos dados moleculares simultaneamente a sofisticadas análises dos dados de campo. Tal modelo, em associação com a análise espacial, é da forma:

$$y = \mu + g + e$$

$$= \mu + g_m + g_{nm} + e$$

$$= \mu + \beta_L x_L + \beta_R x_R + g_{nm} + e$$

$$= Xb + Z\beta_L x_L + Z\beta_R x_R + Zg_{nm} + \xi + \eta \text{ , em que:}$$

$g = g_m + g_{nm}$: efeito genotípico.

$g_m = \beta_L x_L + \beta_R x_R$: efeito genotípico do QTL marcado.

g_{nm}: efeito genético dos QTLs não marcados.

x_L e x_R : informações moleculares (escores para presença ou ausência dos alelos dos marcadores) associadas aos marcadores flanqueadores à esquerda e à direita do QTL, respectivamente, as quais são tratadas como covariáveis.

β_L e β_R: coeficientes de regressão que associam *g* a x_L e x_R , respectivamente.

Este modelo pode ser estendido para incluir também os efeitos de competição e de interação genótipo x ambiente segundo um modelo fator analítico multiplicativo misto (FAMM) (ver Capítulo 8).

Assumindo QTLs como efeitos aleatórios, a significância dos efeitos dos locos marcados pode ser testada por meio do REMLRT no contexto dos modelos lineares mistos. Um modelo incluindo o efeito de QTL é da forma $y = Xb + Zq + Zg + e$, em que *q* é um vetor de efeitos genéticos associados ao QTL marcado, com distribuição $q \sim N(0, G\sigma_q^2)$, em que σ_q^2 é a variância genética do QTL marcado e G é a matriz de covariância para q, condicional à informação do marcador. Para indivíduos não endógamos, G representa a proporção de alelos idênticos por descendência no QTL marcado. Quando se assume que nenhum QTL marcado está segregando na população, o modelo misto é da forma $y = Xb + Zg + e$, o qual é hierárquico ao anterior. Assim, a presença de um QTL em uma particular posição no cromossomo pode ser testada pelo REMLRT envolvendo estes dois modelos. Estes modelos podem ser estendidos pela incorporação de efeitos espaciais, competição e interação genótipo x ambiente.

Um método eficiente de análise de QTL foi apresentado por Gilmour (2007). É denominado mapeamento via regressão sob modelos mistos (MMRM) e é adequado para populações de retrocruzamento e F2. Relaciona-se ao mapeamento por intervalo e por intervalo composto, mas difere no sentido em que se testa a presença de QTL´s em cada grupo de ligação, antes de fazer a regressão. Para isso, o método MMRM inicialmente ajusta todos os marcadores como efeitos

aleatórios com variância comum dentro de cada grupo de ligação. A significância dos efeitos dos marcadores é avaliada via REMLRT e, se existir um componente de variância significativo associado com um grupo de ligação, a análise de QTL via regressão prossegue.

A análise de QTL em plantas perenes e em animais domésticos utiliza métodos estatísticos também muito empregados em genética humana. Os livros de Lange (1997) e Sham (1998) contemplam de forma detalhada a Genética Quantitativa Humana e a Estatística em Genética Humana, sendo referências úteis.

CAPÍTULO 6
ANÁLISE ESTATÍSTICA DA INTERFERÊNCIA ENTRE TRATAMENTOS E COMPETIÇÃO

1 INTERFERÊNCIA ENTRE TRATAMENTOS NOS EXPERIMENTOS DE CAMPO

A análise de experimentos de campo com plantas deve ser baseada em abordagens realísticas, levando-se em consideração o processo biológico associado ao caráter avaliado, bem como as influências ambientais. Existem duas suposições básicas associadas ao modelo clássico de análise de experimentos em blocos. Primeira, que a fertilidade associada com as parcelas no bloco é constante, pelo menos aproximadamente. Segunda, que a resposta em uma parcela, devida a um determinado tratamento, não afeta a resposta em uma parcela vizinha. A primeira suposição está associada a um efeito ambiental ou residual denominado tendência espacial, enquanto a segunda suposição diz respeito a um componente do efeito de tratamento (genético), denominado interferência ou competição. Ajustamentos para estes dois efeitos tendem a reduzir vícios e a melhorar a análise de experimentos de campo.

A correção para os efeitos de tendência espacial tende a aumentar a estimativa da herdabilidade e a precisão na seleção visto que tais efeitos são de natureza ambiental ou residual. Por outro lado, a correção para os efeitos de competição tende a reduzir as estimativas de herdabilidade, visto que a interferência está associada aos efeitos genéticos de tratamentos. Mesmo assim, tal correção aumenta a precisão na avaliação genotípica. A interferência somente pode ocorrer em algumas espécies de plantas e em determinada fase do crescimento. Assim, depende da biologia de cada particular espécie. Tais efeitos têm sido relatados em várias culturas anuais (Talbot et al., 1995) e perenes. Em plantas perenes, a competição tem sido encontrada em espécies florestais (Correll e Anderson, 1983; Resende e Thompson, 2003; Resende et al., 2005), cacau (Glendinning e Vernon, 1965; Lotode e Lachenaud, 1988), dendê (Nouy et al., 1990) e café robusta (Montagnon et al., 2001). A interferência depende também do tamanho e forma das parcelas e tem sido observada em cana-de-açúcar, em ensaios com parcelas de uma só fileira (Stringer e Cullis, 2002a e b).

Uma modalidade de interferência em experimentos de campo refere-se à competição entre genótipos ou variedades, a qual tende a gerar correlações negativas entre a performance de parcelas vizinhas. Neste caso, variedades mais agressivas tendem a ter as suas performances superestimadas nos experimentos de campo, visto que competem com variedades mais sensíveis, as quais têm as suas produtividades subestimadas. Nos plantios comerciais puros, tais variedades agressivas apresentam depressão (devida à competição intra-genotípica) em suas produtividades em relação às performances exibidas no experimento. A inclusão do efeito de competição nos modelos de análise elimina esta distorção. Modelos para avaliação da agressividade e sensibilidade de genótipos foram apresentados por Kempton (1982). Basicamente, dois modelos de competição podem ser empregados. O modelo fenotípico, o qual trata o valor fenotípico dos vizinhos como uma covariável, e o modelo genotípico, o qual divide o efeito de tratamento em dois componentes: efeito direto no próprio tratamento ou genótipo e efeito indireto nos vizinhos (Besag e Kempton, 1986; Stringer e Cullis, 2002; Resende, Stringer, Cullis e Thompson, 2004; 2005). Maiores detalhes são apresentados em tópico seguinte.

Importantes características dos efeitos da interferência entre plantas e da tendência espacial são suas influências nos modelos ajustados. A tendência espacial gera auto-correlações positivas entre parcelas ou planta vizinhas e a interferência entre plantas conduz a autocorrelações negativas entre elas. O ajuste inicial de modelos espaciais pode revelar a necessidade de modelos de competição. Estimativas positivas e altas (superiores a 0,30) de coeficientes de auto-correlação obtidas da análise espacial revela que a tendência espacial é predominante sobre a competição e estimativas negativas ou próximas de zero de coeficientes de auto-correlação revelam a presença de fortes efeitos de competição, provavelmente em conjunto com tendência espacial. Outra alternativa é ajustar inicialmente um modelo de competição, o qual pode revelar a presença de tais efeitos. Em alguns casos, a modelagem de apenas um desses dois efeitos pode ser inapropriada e a modelagem simultânea dos dois pode ser desejável. Durban, Currie e Kempton (2001) relataram estimativas de tendência em fertilidade e de competição mais altas em beterraba açucareira quando esses dois efeitos foram ajustados simultaneamente do que quando foram modelados isoladamente.

2 MODELOS DE INTERFERÊNCIA E COMPETIÇÃO

Vários modelos de competição entre plantas foram propostos. Mead (1967) apresentou a teoria original da competição entre plantas em estandes puros. Outros trabalhos relevantes foram os de Pierce (1957), Draper e Guttman (1980), Kempton (1982), Besag e Kempton (1986), Pithuncharurnlap, Basford e Federer (1993), Talbot et al. (1995), Durban, Hackett e Currie (1999), Durban, Currie e Kempton (2001), Stringer e Cullis (2002b), Resende e Thompson (2003); Resende et al (2004; 2005). Tais modelos serão considerados com detalhes nesse capítulo. Pierce (1957), Draper e Guttman (1980), Kempton (1982), Besag e Kempton (1986) consideraram isoladamente a competição. Pithuncharurnlap, Basford e Federer (1993) consideraram simultaneamente a tendência ambiental espacial por meio do modelo autoregressivo em uma direção de Gleeson e Cullis (1987) e a competição via a abordagem genotípica de Besag e Kempton (1986). Durban, Hackett e Currie (1999) e Durban, Currie e Kempton (2001) consideraram simultaneamente os dois efeitos modelando a tendência via splines cúbicas dentro de blocos e a interferência via o modelo fenotípico de Kempton (1982). Stringer e Cullis (2002b), Resende e Thompson (2003) e Resende et al (2004;

2005) realizaram a modelagem conjunta dos efeitos espaciais e de competição por meio do modelo autoregressivo em duas direções de Gilmour, Cullis e Verbyla (1997) e os modelos fenotípico de Kempton (1982) e genotípico de Besag e Kempton (1986), respectivamente.

Uma forma simples de diagnosticar a presença de efeitos de competição em ensaios de campo consiste em calcular os coeficientes de autocorrelação residual no sentido das linhas e das colunas da grade experimental. Baixos valores (positivos ou negativos) dos coeficientes indicam a presença de competição. Variogramas exibindo picos, altos e baixos pontos (alternando cumeeiras) também revelam correlação negativa entre resíduos e, portanto, competição.

2.1 Modelo Fenotípico de Interferência

Kempton (1982) apresentou o seguinte modelo para competição:

$$Y_{ij} = \tau_i + \beta X_j + \varepsilon_{ij}, \quad (1)$$

em que:

y_{ij} : valor observado do genótipo i na parcela j;

τ_i : efeito fixo do tratamento ou genótipo i;

β : coeficiente de competição, comum a todos os genótipos;

X_j : média das parcelas vizinhas do genótipo i na parcela j;

ε_{ij} : erro independente e com distribuição normal com media zero e variância σ^2.

O modelo assume observações ajustadas para a média geral e ignora os efeitos de bloco. A covariável X é dada por $X = \sum y / p$, em que p é o número de parcelas vizinhas consideradas. Normalmente p pode ser 2 (avaliação em nível de parcela, várias plantas por parcela), 4 (avaliação em nível de plantas, com uma ou várias plantas por parcela) ou 8 (avaliação em nível de plantas, com uma ou várias plantas por parcela).

O efeito τ_i representa o efeito esperado do genótipo quando a variedade é cultivada sob o estresse competitivo do ensaio. Sua performance em monocultura é estimada por $\tau_{ic} = \tau_i /(1-\beta)$. Uma vez que β é negativo, pode-se depreender que as performances dos melhores tratamentos são reduzidas após a correção para competição. Isto é devido ao fato de que, sob competição, as variedades mais agressivas têm suas produtividades superestimadas em detrimento das variedades mais sensíveis. Se o experimentador está interessado em avaliar a performance varietal em monocultivos, esta correção deve ser feita. As diferenças observadas entre a performance dos genótipos nos ensaios e em plantios comerciais surgem em parte devido ao fato de que a alocação das variedades nos ensaios não é balanceada quanto às variedades vizinhas, mas principalmente porque uma variedade selecionada tende a ser muito competitiva no ensaio e depois exibe uma natural depressão em produtividade quando plantada em monocultivo. Isto tem sido observado por exemplo em cana-de-açúcar, revelando a necessidade de correções.

Os parâmetros podem ser estimados simultaneamente por quadrados mínimos por meio do seguinte conjunto de equações:

$$\hat{\tau}_i = (g/n)\sum_j (Y_{ij} - \hat{\beta} X_j)$$

$$\hat{\beta} = \sum_j^n (Y_{ij} - \hat{\tau}_i) X_j /(\sum_{j=1}^n X_j^2) \cdot$$

O somatório na equação para τ_i estende somente sobre o conjunto de parcelas j contendo o genótipo i (n/g parcelas, em que n é o número total de parcelas no ensaio e g é o número de genótipos ou tratamentos). Na equação para β, todas as parcelas são usadas, uma vez que o coeficiente de competição é comum a todos os genótipos. β é um coeficiente de regressão que relaciona os resíduos com o valor médio (como uma covariável) das plantas ou parcelas vizinhas.

Esta abordagem de quadrados mínimos é adequada quando a covariável é uma variável diferente daquela associada ao caráter de interesse. Por exemplo, é válida quando o caráter de interesse é a produção e a covariável é altura das plantas vizinhas. Entretanto, quando a covariável é definida como a mesma variável de interesse (por exemplo, ambos sendo a produção), a abordagem de quadrados mínimos produz uma estimativa inválida de β, uma vez que o coeficiente

de competição aparece em ambos, na média e na variância de y. Nesse caso, uma estimação eficiente pode ser realizada via máxima verossimilhança. A significância de β no modelo pode ser testada via o teste da razão de verossimilhança. A omissão do efeito de competição pode aumentar a deviance ℓ do modelo. Para testar a significância da competição, após o ajuste para os efeitos de variedades, a estatística $\ell(y\,|\,\hat{\sigma},\hat{\tau}) - \ell(y\,|\,\hat{\sigma},\hat{\tau},\hat{\beta})$ deve ser usada, a qual, sob a hipótese de nulidade, é aproximada por uma distribuição χ^2 com um grau de liberdade.

O mesmo modelo pode ser rescrito pela consideração de somente duas parcelas ou plantas como vizinhas:

$$Y_{ij} = \tau_i + (1/2)\beta(Y_{j+1,s} + Y_{j-1,t}) + \varepsilon_{ij} \qquad (2)$$

em que $Y_{j+1,s}$ e $Y_{j-1,t}$ são as performances (para o mesmo caráter) dos genótipos s e t em parcelas vizinhas ao genótipo i na parcela j. Em algumas situações, os coeficientes de competição dependem dos específicos genótipos cultivados nas parcelas. Em tais casos, coeficientes de competição específicos $\beta_{is} = \delta_i\gamma_s$ podem ser demandados, em que δ_i representa a sensibilidade do genótipo i à competição e γ_s representa a agressividade do genótipo s e pode ser padronizado de forma que $\sum_s \gamma_s = g$. Então o modelo (2) pode ser rescrito como $Y_{ij} = \tau_i + (1/2)(\beta_{is}Y_s + \beta_{it}Y_t) + \varepsilon_{ij}$.

Em notação matricial, o modelo (2) pode ser rescrito como (Besag e Kempton, 1986):

$$y = Xb + Z\tau + \beta Wy + \varepsilon \qquad (3)$$

em que:

Z: matriz de incidência dos efeitos de tratamentos ou genótipos.

W: é uma matriz regressora ou de pesos, n x n, a qual tem os elementos fora da diagonal (j, j $\pm$ 1) ou das diagonais secundárias iguais a (1/2) e os demais elementos igual a zero.

b: é o vetor de efeitos associados ao delineamento experimental tais como blocos, com matriz de incidência X.

O vetor τ pode ser interpretado como efeitos genotípicos centrados, na ausência de competição ou sob o estresse competitivo médio do ensaio. Mas quando plantadas em monocultivo,

as melhores variedades produzirão um ambiente mais competitivo do que aquele médio do ensaio e então não terão performance tão boa quanto nos ensaios. Então, τ deve ser dividido por um fator $(1-\beta)$ para representar a performance em estandes puros. Os efeitos de competição aumentam a amplitude e a variabilidade da distribuição de todos os efeitos de genótipos, uma vez que amplificam os valores dos genótipos mais agressivos e diminuem os efeitos dos genótipos mais sensíveis. A correção usando o fator $(1-\beta)$ causa encolhimento ou regressão (*shrinkage*) nos efeitos genotípicos, conduzindo a resultados mais realísticos.

De acordo com Kempton (1985), uma forma alternativa para o modelo (2) é $Y_{ij} = \tau_{ic} + \beta(Y_{j+1} + Y_{j-1} - 2Y_j) + \varepsilon_{ij}$. Nesse caso, os efeitos de tratamentos são ajustados, já corrigidos para os efeitos de vizinhança, isto é, já representam a produtividade em estandes puros.

2.2 Modelo Genotípico de Interferência

Draper e Guttman (1980) ignoraram os erros em Y_s e Y_t e usaram o modelo (2) como

$$Y_{ij} = \tau_i + (1/2)\beta(\tau_s + \tau_t) + \varepsilon_{ij} \qquad (4)$$

Esse modelo considera que a competição é mais determinada pelo genótipo do que pelo fenótipo das plantas. Isto faz sentido, uma vez a agressividade e a sensibilidade dos genótipos têm uma base genética, ou seja, apresentam controle genético e dependem também de outros caracteres na planta tais como altura, tamanho do dossel e capacidade de perfilhamento e enraizamento, dentre outros. Em tal modelo, o coeficiente de regressão relaciona os efeitos genéticos dos vizinhos ao valor residual da planta ou parcela central.

Pierce (1957) considerou um modelo de interferência de parcela no qual cada tratamento i tem um efeito direto τ_i na parcela na qual é aplicado e um efeito indireto ϕ_i na parcela vizinha. Esse modelo é genérico e aplica-se não somente a tratamentos genéticos, mas a outros tratamentos fitotécnicos também, como fertilizantes, defensivos, etc. A competição genotípica pode ser considerada sob a ótica desse modelo, uma vez que as causas (luminosidade, competição radicular,

etc) exatas da competição são desconhecidas. De acordo com Besag e Kempton (1986), esse modelo é da forma:

$$y = Xb + Z\tau + NZ\phi + \varepsilon \qquad (5)$$

em que:

ϕ : vetor dos efeitos centrados de tratamentos (genótipos) sobre os vizinhos (efeitos indiretos ou associativos), os quais são genotípicos e não fenotípicos.

N: matriz de incidência de vizinhança, de dimensão n x n, composta por 0 e 1.

Pode ser visto explicitamente no modelo (5) que os efeitos de competição referem-se a efeitos de tratamentos (dependem da matriz Z) e não a efeitos residuais. Devido a essa razão, somente a abordagem auto-regressiva para os resíduos pode ser inapropriada para contemplar a competição entre plantas ou entre parcelas.

Draper e Guttman (1980) incluíram um caso especial de (5) no qual $\phi_i = \lambda\tau_i$, em que λ é um coeficiente de interferência, comum a todos os genótipos. O modelo é:

$$\begin{aligned} y &= Xb + H\tau + \varepsilon \\ &= Xb + Z\tau + NZ\lambda\tau + \varepsilon \end{aligned} \qquad (6)$$

em que $H = (I + \lambda N)Z$, de forma que o modelo é não linear em τ e λ. Os efeitos de tratamentos em estandes puros é dado por $\tau_i^* = (1 + v\lambda)\tau_i$.

O componente ϕ_i em (5) pode ser positivo ou negativo, dependendo da agressividade do tratamento. Se negativo (para variedades agressivas), o valor absoluto de ϕ_i deve ser subtraído de τ_i por meio de $\tau_i^* = \tau_i + v\phi_i$, propiciando os efeitos de tratamentos para plantios puros, em que v é o número de vizinhos considerados. Se positivo (variedades sensíveis), ϕ_i será somado na expressão para τ_i^*.

O efeito de vizinhança não é sempre correlacionado (negativamente) com o caráter em avaliação, uma vez que pode depender de outros caracteres, como altura e vigor das plantas. Nos casos em que ϕ_i não tem relação com τ_i, os modelos (1), (2), (3), (4) e (6) são inadequados porque

consideram um único coeficiente de competição para todos os genótipos. Então, o modelo (5) tende a ser melhor, uma vez que permite que os efeitos genotípicos de competição sejam individualmente especificados. Também, um ordenamento baseado no componente ϕ_i pode ser realizado visando à seleção de genótipos com alta produção e baixa agressividade competitiva para uso em plantios adensados.

2.3 Modelagem Conjunta da Interferência e Tendência em Fertilidade

Pithuncharurnlap, Basford e Federer (1993) incluíram competição e tendência em um modelo espacial. Usaram o modelo auto-regressivo em uma dimensão de Gleeson e Cullis (1987) para modelagem da tendência em fertilidade e o modelo de competição genotípica (5) de Besag e Kempton (1986) para modelagem da interferência. O modelo é da forma:

$$y = Xb + Z\tau + NZ\phi + \xi + \eta \qquad (7)$$

em que:

ξ : vetor aleatório de erros correlacionados.

η : vetor aleatório de erros não correlacionados.

A competição foi modelada como parte da estrutura de tratamentos e a tendência em uma dimensão foi modelada como parte da estrutura dos erros.

Durban, Currie e Kempton (2001) comentaram a respeito do problema estatístico de modelar simultaneamente dois tipos de correlação local. De acordo com eles, dificuldades existem na modelagem conjunta da tendência e competição pelo modelo fenotípico, uma vez que ambos são efeitos de correlação. Então sugeriram diferentes mecanismos para especificar os dois efeitos, permitindo que sejam estimados separadamente.

Durban, Hackett e Currie (1999) e Durban, Currie e Kempton (2001) consideraram simultaneamente os dois efeitos, modelando a tendência especial por meio de splines cúbicas dentro de blocos e a interferência pelo modelo fenotípico de Kempton (1982). Entretanto, splines

podem não ser a melhor opção para modelagem da tendência espacial. Em geral, os modelos com estrutura autoregressiva separável em duas dimensões propiciam um melhor ajuste (Cullis e Gleeson, 1991; Gilmour, Cullis e Verbyla, 1997).

Stringer e Cullis (2002b), Resende e Thompson (2003) e Resende et al. (2004; 2005) usaram o mesmo modelo descrito em (7), mas assumiram τ_i e ϕ_i como efeitos aleatórios (aqui denominado modelo 8). Nesse caso, existe uma covariância entre τ_i e ϕ_i. A matriz de covariância entre eles equivale a:

$G = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$, em que $g_{\tau\tau}$ é o componente de variância para os efeitos genotípicos diretos, $g_{\phi\phi}$ é o componente de variância associado aos efeitos genotípicos indiretos sobre os vizinhos e $g_{\tau\phi}$ é a covariância entre os efeitos diretos na própria planta e indiretos sobre os vizinhos.

Recentemente um modelo idêntico ao (8) passou a ser usado também no melhoramento animal (Van Vleck e Cassady, 2005; Muir, 2005; Arango et al., 2005, Cappa e Cantet, 2006).

De acordo com o modelo (6) de Draper e Guttman (1980), a matriz de variância-covariância G é dada por:

$$G = \begin{pmatrix} g_{\tau\tau} & \lambda_1 g_{\tau\tau} \\ \lambda_1 g_{\tau\tau} & \lambda_1^2 g_{\tau\tau} \end{pmatrix}.$$

Essa matriz de covariância é de posto reduzido (posto = 1). Thompson et al. (2003) descrevem como lidar com modelos desse tipo.

Os modelos de competição aplicados em plantas perenes e espécies florestais têm sido os mesmos (com pequenas modificações) aplicados em culturas anuais. Correll e Anderson (1983) aplicaram o modelo de competição de Draper e Guttman (1980), seguido pela análise espacial, segundo o método de Papadakis para considerar, de maneira não simultânea, a interferência e a tendência espacial, respectivamente. Magnussen e Yeatman (1987) usaram duas abordagens: o índice de competição de Hegyi (1974) como covariável e uma modificação do modelo de competição de Kempton (1982). O índice de competição de Hegyi (1974) foi proposto no contexto da pressão

competitiva em árvores individuais em estandes naturais e é dado por: $C_i = (\sum_{j=1}^{8} Y_j / Y_i) / Dist_{ij}$, em que:

C_i : índice de competição da planta alvo i.

Y_i : valor observado da planta alvo i.

Y_j : valor observado da planta competidora j.

$Dist_{ij}$: distância entre as plantas i e j.

O uso desse índice como uma covariável produz resultados similares àqueles obtidos com o método fenotípico de Kempton (1982) quando aplicado à média dos oito vizinhos e assumindo iguais espaçamentos entre eles e a árvore alvo. Então, a vantagem do índice de Hegyi refere-se apenas à consideração das diferentes distâncias entre a árvore alvo e as vizinhas. Leonardecz-Neto (2002) também aplicou esse índice em espécies florestais.

A modificação no método de Kempton (1982), introduzida por Magnussen e Yeatman (1987), foi a consideração de dois coeficientes de competição β , um para indivíduos de diferentes tratamentos e outro para indivíduos do mesmo tratamento em uma parcela, ou seja, um coeficiente de competição para indivíduos aparentados e outro para não aparentados. Esse tipo de modelo é um esquema auto-normal de primeira ordem de um processo de Markov bidimensional (Besag, 1974) desde que o delineamento experimental seja considerado como um latice regular de pontos com variáveis contínuas tendo distribuição normal multivariada, e assumindo estabilidade no tempo e espaço.

Magnussen (1994) considerou o simultâneo ajustamento para efeitos espaciais e de competição por meio do uso de modificações da abordagem de Correll e Anderson (1983), baseada no método de Papadakis. Kusnandar (2001) estendeu o modelo Kempton (1982) para duas dimensões, considerando a competição nas direções das linhas e colunas, via um modelo misto. Montagnon et al. (2001) relataram o primeiro trabalho lidando com competição em cafeeiros. Eles usaram coeficientes de competição específicos para cada tratamento, porém apenas em nível dos resíduos. Para isso, usaram a técnica de regressão linear múltipla para estimar os efeitos de competição ou efeitos compartilhados, conforme Gallais (1975). Também Gallais (1989, página 79)

descreve interessantes modelos fatoriais para avaliação e seleção envolvendo experimentação e plantios puros e em mistura, aplicados a plantas forrageiras. A abordagem considera, sob a ótica de modelos de competição, as combinações: experimentação em plantio puro – plantio comercial puro; experimentação em plantio puro – plantio comercial em mistura; experimentação em plantio misto – plantio comercial puro e experimentação em plantio misto – plantio comercial em mistura.

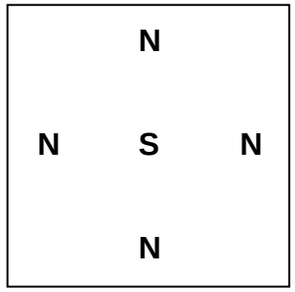
Os modelos de competição usados em espécies perenes relatados anteriormente ou não consideraram efeitos de competição específicos para cada genótipo ou não particionaram os efeitos de genótipos em efeitos diretos no próprio genótipo e indiretos nos vizinhos. Também para a modelagem conjunta da tendência e da competição, a abordagem usada foi a de Papadakis a qual pode não ser a melhor, ou um processo auto-regressivo em uma só dimensão. Além disso, o uso do modelo fenotípico de interferência e do índice de Hegyi, nos quais a covariável é definida como sendo a mesma variável referente ao caráter de interesse (por exemplo, ambos sendo a altura das plantas), a abordagem de quadrados mínimos conduz a uma estimativa inválida de β. Uma estimação eficiente pode ser realizada usando máxima verossimilhança perfilada (ver Capítulo 4 e tópico 3, a seguir), técnica essa não usada nos trabalhos relatados anteriormente. Baseando-se nesses antecedentes, Resende e Thompson (2003) propuseram uma nova modelagem simultânea para tendência e efeitos de competição em plantas perenes, a qual considera esses aspectos mencionados.

3 MODELAGEM DA COMPETIÇÃO EM PLANTAS PERENES E SEMI-PERENES

3.1 Modelos com Efeitos Genéticos Diretos no Próprio Genótipo e Indiretos nos Vizinhos

O modelo (8) apresentado anteriormente pode ser usado para qualquer número de vizinhos considerados. Assumindo iguais espaçamentos entre a planta alvo (S) e as plantas vizinhas (N), têm-se os seguintes aspectos associados à modelagem da competição, conforme Resende e Thompson (2003):

Caso de 4 Competidores: Distribuição dos Vizinhos



Os efeitos de competição nos quatro vizinhos podem ser especificados individualmente (quando o efeito de vizinhança depende do sombreamento, por exemplo) como vizinhos a leste, oeste, norte e sul, ou em conjunto gerando um só coeficiente de competição ϕ_{HV} englobando tanto vizinhos horizontais quanto verticais. Esse último tipo de modelo é detalhado abaixo.

Modelo

$$y = Xb + Z\tau + N_{HV} Z\phi_{HV} + \xi + \eta \quad (8)$$

τ : vetor aleatório de efeitos genotípicos em ausência de competição ou sob o estresse competitivo médio do ensaio.

N_{HV} : matriz de incidência para vizinhos horizontais e verticais.

ϕ_{HV} : vetor aleatório de efeitos genotípicos indiretos sobre vizinhos horizontais e verticais.

ξ : vetor aleatório de erros correlacionados, modelados por processo auto-regressivo em duas dimensões.

O ajuste desse modelo é similar ao procedimento adotado para modelos com efeitos maternos usados no melhoramento animal. Um exemplo detalhado é apresentado em Resende e Rosa-Perez (2001). No entanto, no presente caso, vários vizinhos são considerados simultaneamente. Isso difere do caso de efeitos maternos, em que um só animal por vez é considerado.

Arranjo de Campo e Matriz de Incidência de Vizinhança (N_{HV})

```
1  5  9
2  6  10
3  7  11
4  8  12

1  2  3  4  5  6  7  8  9  10  11  12

0  1  0  0  1  0  0  0  0  0  0  0
   0  1  0  0  1  0  0  0  0  0  0
      0  1  0  0  1  0  0  0  0  0
         0  0  0  0  1  0  0  0  0
            0  1  0  0  1  0  0  0
               0  1  0  0  1  0  0
                  0  1  0  0  1  0
                     0  1  0  0  1
                        0  0  0  0
                           0  1  0
                              0  1
                                 0
```

Nos casos em que Z é uma matriz identidade, tem-se que $N_{HV} Z = N_{HV}$, ou seja, a matriz de incidência dos efeitos genotípicos indiretos nos vizinhos equivale à própria matriz de incidência de vizinhos N_{HV}. Matriz de incidência de vizinhos idêntica foi também apresentada recentemente por Muir (2005).

Matriz de Covariância dos Efeitos Genotípicos Diretos e Vicinais

$$G = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi_{HV}} \\ g_{\tau\phi_{HV}} & g_{\phi_{HV}\phi_{HV}} \end{pmatrix}$$

Efeitos Genotípicos Corrigidos

$\tau_i^* = \tau_i + v_{HVi}\phi_{HV_i}$

v_{HVi}: somatório do número de vizinhos horizontais e verticais do genótipo i, isto é, 4.

Caso de 8 Competidores: Distribuição dos Vizinhos

N	**N**	**N**
N	**S**	**N**
N	**N**	**N**

Levando-se em conta as diferentes distâncias entre a árvore alvo (S) e as vizinhas (N), o modelo (8) deve ser estendido para o modelo (9).

Modelo

$$y = Xb + Z\tau + N_{HV}Z\phi_{HV} + N_D Z\phi_D + \xi + \eta \quad (9)$$

τ: vetor aleatório de efeitos genotípicos em ausência de competição ou sob o estresse competitivo médio do ensaio;

N_{HV}: matriz de incidência para vizinhos horizontais e verticais.;

ϕ_{HV}: vetor aleatório de efeitos genotípicos indiretos sobre vizinhos horizontais e verticais;

N_D: matriz de incidência para os vizinhos diagonais;

ϕ_D: vetor aleatório de efeitos genotípicos indiretos sobre vizinhos diagonais.

Arranjo de Campo e Matriz de Incidência de Vizinhança Diagonal (N_D)

```
1  5  9
2  6  10
3  7  11
4  8  12

1  2  3  4  5  6  7  8  9  10  11  12

0  0  0  0  0  1  0  0  0  0  0  0
   0  0  0  1  0  1  0  0  0  0  0
      0  0  0  1  0  1  0  0  0  0
         0  0  0  1  0  0  0  0  0
            0  0  0  0  0  1  0  0
               0  0  0  1  0  1  0
                  0  0  0  1  0  1
                     0  0  0  1  0
                        0  0  0  0
                           0  0  0
                              0  0
                                 0
```

Matriz de Covariância dos Efeitos Genotípicos Diretos e Vicinais

$$G = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi_{HV}} & g_{\tau\phi_D} \\ & g_{\phi_{HV}\phi_{HV}} & g_{\phi_{HV}\phi_D} \\ & & g_{\phi_D\phi_D} \end{pmatrix}$$

Efeitos Genotípicos Corrigidos

$$\tau_i^* = \tau_i + v_{HVi}\phi_{HV_i} + v_{Di}\phi_{D_i}$$

v_{HVi}: somatório do número de vizinhos horizontais e verticais do genótipo i, isto é, 4;

v_{Di}: número de vizinhos diagonais do genótipo i, isto é, 4.

Caso de 8 Competidores: Distribuição dos Vizinhos

NW	**N**	**NE**
W	**S**	**E**
SW	**S**	**SE**

Os efeitos de competição em 8 vizinhos podem ser especificados individualmente (quando esses efeitos dependem do sombreamento, tal qual como ocorre em espécies florestais) como vizinhos a norte (N), sul (S), leste (E), oeste (W), nordeste (NE), sudeste (SE), noroeste (NW) e sudoeste (SW). Um modelo generalizado adequado para esse caso, em qualquer arranjo experimental, é dado pelo modelo (10), conforme Resende e Thompson (2003).

Modelo

$$y = Xb + Z\tau + N_E Z\phi_E + N_W Z\phi_W + N_N Z\phi_N + N_S Z\phi_S + N_{NE} Z\phi_{NE} + N_{SE} Z\phi_{SE} + N_{NW} Z\phi_{NW} + N_{SW} Z\phi_{SW} + \xi + \eta \quad (10)$$

em que:

τ : vetor aleatório de efeitos genotípicos em ausência de competição ou sob o estresse competitivo médio do ensaio;

N_E : matriz de incidência para vizinhos a leste;

ϕ_E : vetor de efeitos genotípicos sobre os vizinhos a leste;

N_W : matriz de incidência para vizinhos a oeste;

ϕ_W : vetor de efeitos genotípicos sobre os vizinhos a oeste;

N_N : matriz de incidência para vizinhos a norte;

ϕ_N : vetor de efeitos genotípicos sobre os vizinhos a norte;

N_S : matriz de incidência para vizinhos ao sul;

ϕ_S : vetor de efeitos genotípicos sobre os vizinhos ao sul;

N_{NE} : matriz de incidência para vizinhos a nordeste;

ϕ_{NE} : vetor de efeitos genotípicos sobre os vizinhos a nordeste;

N_{SE} : matriz de incidência para vizinhos a sudeste;

ϕ_{SE} : vetor de efeitos genotípicos sobre os vizinhos a sudeste;

N_{NW} : matriz de incidência para vizinhos a noroeste;

ϕ_{NW} : vetor de efeitos genotípicos sobre os vizinhos a noroeste;

N_{SW} : matriz de incidência para vizinhos a sudoeste;

ϕ_{SW} : vetor de efeitos genotípicos sobre os vizinhos a sudoeste.

Este modelo completo bem como modelos aninhados dentro desse podem ser usados para inferir sobre a significância de posições de vizinhança específicas. O modelo final selecionado deve permitir a covariância entre os efeitos aleatórios de vizinhança retidos. Em espécies florestais, Resende e Thompson (2003) encontraram significância para apenas uma dessas posições, de acordo com o sentido de caminhamento do sol, e um modelo com um só efeito indireto nos vizinhos foi ajustado.

Após a determinação da posição do vizinho significativo, tal modelo simultâneo é dado por $y = Xb + Z\tau + NZ\phi + \xi + \eta$, em que τ e ϕ são os efeitos diretos no tratamento e indiretos nos vizinhos, assumidos como aleatórios e com matriz de covariância $G = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$, em que $g_{\tau\tau}$ é a variância associada aos efeitos genotípicos diretos e $g_{\phi\phi}$ é a variância associada aos efeitos genotípicos indiretos de competição, sendo o numerador da herdabilidade dos efeitos de competição. O parâmetro $g_{\tau\phi}$ é covariância entre os efeitos diretos e indiretos e é o numerador da correlação

genética entre a produtividade e a agressividade das variedades, dada por $r_{\tau\phi} = g_{\tau\phi} / (g_{\tau\tau} g_{\phi\phi})^{1/2}$. Esta correlação é, em geral, negativa, evidenciando que as melhores variedades são beneficiadas nos experimentos. A seleção deve então ser baseada em $\hat{\tau} + \hat{\phi}$, em que $\hat{\phi}$ é negativo nas variedades mais agressivas. A seleção pode basear-se também em $\hat{\phi}$ visando a identificação de genótipos adequados a plantios adensados como, por exemplo, no melhoramento do cafeeiro e do dendezeiro.

Efeitos Genotípicos Corrigidos Associados ao Modelo Completo

$$\tau_i^* = \tau_i + \phi_E + \phi_W + \phi_N + \phi_S + \phi_{NE} + \phi_{SE} + \phi_{NW} + \phi_{SW}$$

Esse modelo generalizado demanda grande quantidade de dados para que possa ser ajustado, uma vez que um grande número de graus de liberdade é necessário para ajustar todos esses efeitos. Ensaios com grande número de tratamentos e limitado número de repetições não são adequados para que se ajuste esse modelo e provavelmente nem os modelos (8) e (9). Nesse caso, uma alternativa é o uso do modelo fenotípico de Kempton (1982) em conjunto com a análise espacial de Cullis e Gleeson (1991). Essa abordagem foi usada por Resende e Thompson (2003) em algumas situações e é descrita no tópico a seguir.

3.2 Modelo de Competição Fenotípica Ajustado via Máxima Verossimilhança Perfilada

O modelo de competição fenotípica mais análise espacial via processo auto-regressivo é dado por:

$$y = Xb + Z\tau + \beta Wy + \xi + \eta \qquad (11)$$

em que:

W: matriz regressora ou de pesos, n x n, a qual, em conjunto com y, propicia o valor médio dos vizinhos como covariável. Em geral, a média de dois ou quatro vizinhos podem ser usadas.

Vizinhos diagonais têm menor probabilidade de afetar as árvores alvo porque apresentam maior distância em relação a elas e também porque produzem efeitos no crescimento das vizinhas mais próximas da árvore alvo. Então, se uma planta diagonal cresce mais que a vizinha mais próxima da árvore alvo, essa será beneficiada pelo menor crescimento da vizinha e não será prejudicada tanto pelo maior crescimento da árvore na posição diagonal. A estimação e a predição sob esse modelo demandam o uso de verossimilhança perfilada, a qual é detalhada a seguir.

Não é possível usar o método REML ordinário para o modelo fenotípico de competição de Kempton (1982), uma vez que o coeficiente de competição aparece em ambos, na média e variância de y. Entretanto, uma generalização do REML pode ser aplicada para estimação dos parâmetros do modelo. Essa generalização envolve o ajustamento da verossimilhança perfilada (por meio do escore perfilado ajustado) para o parâmetro de interesse em uma classe geral de modelos. Tal ajustamento pode ser feito pelo método de McCullagh e Tibshirani (1990), o qual remove o vício das estimativas de máxima verossimilhança.

A inferência na presença de parâmetros de *nuisance* é um problema difícil em estatística. Sob a perspectiva da verossimilhança, a abordagem mais simples refere-se à eliminação (via maximização) dos referidos parâmetros para valores fixos dos parâmetros de interesse e então construir o que é denominado verossimilhança perfilada. Em outras palavras, tal solução refere-se à substituição dos parâmetros de *nuisance* na função de verossimilhança por suas estimativas de máxima verossimilhança obtidas sob valores fixados dos parâmetros de interesse. Isto produz a verossimilhança perfilada. Essa é então tratada como uma função de verossimilhança ordinária para estimação e inferência sobre os parâmetros de interesse. Infelizmente, com grande número de parâmetros de *nuisance*, esse procedimento pode produzir estimativas ineficientes e inconsistentes. Os problemas inerentes ao uso de verossimilhanças perfiladas são a geração de estimativas viciadas dos parâmetros e otimistas dos desvios padrões.

Modificações na verossimilhança perfilada com o objetivo de aliviar esses problemas foram propostas. Barndorff-Nielsen (1983, 1986) propuseram a verossimilhança perfilada modificada, a qual é intimamente relacionada à verossimilhança perfilada condicional proposta por Cox e Reid (1987) na qual é sugerido um teste de razão de verossimilhança construído a partir da distribuição condicional das observações dadas as estimativas de máxima verossimilhança dos parâmetros de *nuisance*.

McCullagh e Tibshirani (1990) propuseram uma abordagem alternativa mais simples denominada verossimilhança perfilada ajustada. Esse método depende da observação de que a função escore computada da função de verossimilhança completa tem: (i) esperança zero; (ii) variância igual ao negativo da matriz derivativa esperada. Uma função escore que tem a propriedade (i) é dita não viesada, enquanto a propriedade (ii) é dita informação não viesada. Por associação, pode ser dito que a função de verossimilhança é não viesada/informação não viesada se sua função escore é não viesada/informação não viesada. Em contraste com a função escore computada do logaritmo da verossimilhança completa, a função escore computada do logaritmo da verossimilhança perfilada não é, em geral, nem não viesada e nem informação não viesada. A idéia de McCullagh e Tibishirani é que a função escore do logaritmo da verossimilhança perfilada pode ser centrada e escalada de forma que ela seja também não viesada e informação não viesada (Durban e Currie, 2000).

McCullagh e Tibshirani (1990) concentraram em dar fórmulas assintóticas para as correções em um contexto bem genérico. Durban e Currie (2000) forneceram expressões exatas para os ajustamentos para o caso de um modelo geral de regressão não linear normal. Em sua forma mais geral, o modelo permite que ambos, a média e a variância de y, dependam do parâmetro de interesse. Um exemplo dessa forma geral é um modelo de regressão com termos auto-regressivos tais quais no modelo fenotípico de competição. O ajustamento exato para a verossimilhança perfilada de tal modelo melhora a estimação dos parâmetros de variância e de competição.

De acordo com o modelo fenotípico de competição, $y = Xb + Z\tau + \beta Wy + \varepsilon$, pode-se escrever:

$Dy = Xb + Z\tau + \varepsilon$, em que $D = I - \beta W$.

$y = D^{-1}Xb + D^{-1}Z\tau + D^{-1}\varepsilon$, onde:

$y \sim N(D^{-1}Xb, \sigma^2 D^{-1}VD^{-1\prime})$.

$\tau \sim N(0, \sigma^2 G)$

$\varepsilon \sim N(0, \sigma^2 R)$.

$V = ZGZ' + R$.

Considerando *b* como um parâmetro de *nuisance* e $\theta = (\beta, \tau, \sigma^2)$ como parâmetros de interesse, o logaritmo da verossimilhança de $Dy \sim N(Xb, \sigma^2 V)$ é dado por: $\ell = \ell(\theta, b; y) = -(n/2)\log 2\pi - (n/2)\log\sigma^2 - (1/2)\log\left|D^{-1}VD^{-1\prime}\right| - (1/2\sigma^2)(Dy - Xb)'V^{-1}(Dy - Xb)$.

A derivada dessa função com respeito a b e igualada a zero fornece a estimativa de máxima verossimilhança de b, a qual é dada por: $\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}Dy$

De acordo com McCullagh e Tibshirani (1990), o logaritmo da verossimilhança perfilada é obtido pela substituição dos parâmetros de *nuisance* por suas estimativas de máxima verossimilhança. Substituindo-se $\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}Dy$ em $\ell = \ell(\theta, b; y)$ produz o logaritmo da verossimilhança perfilada ℓ_p, a qual, ignorando termos constantes, é equivalente a :

$$\ell_p(\theta; y) = -(n/2)\log\sigma^2 - (1/2)\log\left|D^{-1}VD^{-1\prime}\right| - (1/2\sigma^2)y'D'PVPDy$$
$$= \log\left|D\right| - (n/2)\log\sigma^2 - (1/2)\log\left|V\right| - (1/2\sigma^2)y'D'PDy$$

em que $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$.

Desse logaritmo da verossimilhança perfilada, as equações para os escores perfilados ajustados são obtidas. Para os parâmetros de variância, essas equações equivalem às equações escores do REML baseado em *Dy*.

O logaritmo da função de verossimilhança residual baseada em *Dy* é dado por:

$$\ell_{\mathrm{Re}} = -[(n-p)/2]\log\sigma^2 + \log\left|D\right| - (1/2)\log\left|V\right| - (1/2)\log\left|X'V^{-1}X\right| - (1/2\sigma^2)y'D'PDy$$

Uma diferença fundamental entre essa função e o logaritmo da função de verossimilhança residual baseada em y é o termo adicional $\log|D|$. Então, o REML baseado Dy pode ser obtido usando os algoritmos padrões usados em *softwares* como o AsReml, Genstat e Selegen-Reml/Blup, mas $\log|D|$ deve também ser obtido e somado em Log L.

A presença do coeficiente de competição (parâmetro de interesse) em ambos, média e variância de y, conduz a dificuldades, mas os ajustamentos do método de McCullagh e Tibishirani se aplicam bem nessa situação e a verossimilhança perfilada ajustada resultante é equivalente ao

REML de Patterson e Thompson (1971). As equações para os escores perfilados ajustados são equivalentes às equações escores do REML baseado na resposta ajustada *Dy*. Na prática, D é substituída por sua estimativa. Os escores ajustados produzem ambos, um ajustamento do tipo REML para as estimativas de componentes de variância e um ajustamento para a estimativa de β, removendo o seu vício. Verossimilhanças perfiladas e escores perfilados ajustados para os parâmetros de interesse são apresentados por Durban e Currie (2000) para o caso de um modelo de efeitos fixos.

No contexto do modelo (11), o parâmetro de competição e os componentes de variância podem ser estimados como relatado a seguir: (i) obtenção de REML (Log L) baseado em Dy para vários valores dados de β ; (ii) obtenção de (Log¦D¦) para vários valores de β; (iii) obtenção do ponto de máximo da função de verossimilhança perfilada ajustada (LogL+Log¦D¦) para vários valores de β.

3.3 Consideração dos Efeitos de Plantas Perdidas

Para inferência sobre tratamentos ou genótipos, os efeitos de plantas perdidas são contemplados pela omissão (ou ajuste como efeito fixo) dos zeros correspondentes às parcelas ou plantas perdidas na coluna referente aos efeitos genéticos diretos e pela consideração dos zeros na coluna referente aos efeitos genéticos indiretos sobre os vizinhos, segundo os modelos 8, 9 e 10. Isto pode ser conseguido por meio da codificação de todos os zeros da coluna de vizinhos como pertencentes a um tratamento ainda não codificado na coluna de tratamentos, isto é, codificando-os como um novo tratamento adicional na coluna de vizinhos. Dessa forma, os efeitos das falhas serão refletidos nas estimativas de ϕ e então também em τ_i^*.

4 APLICAÇÕES EM EXPERIMENTOS COM CANA-DE-AÇÚCAR, EUCALIPTO E PÍNUS

4.1 Modelo Fenotípico de Competição via Verossimilhança Perfilada em Cana-de- Açúcar

A seguir, são apresentados resultados associados à avaliação do caráter número de colmos de 128 clones de cana-de-açúcar, no delineamento de blocos ao acaso com duas repetições. Esses dados foram gentilmente cedidos pelo professor Dr. Márcio Henrique Pereira Barbosa, da Universidade Federal de Viçosa. Somente o modelo fenotípico de competição pode ser aplicado em virtude do baixo número de repetições empregado, fato que dificulta o ajuste de modelos mais complexos em função do insuficiente número de graus de liberdade. Os resultados são apresentados nas Tabelas 15, 16 e 17.

Tabela 15. Máximo do logaritmo da função de verossimilhança residual (Log L) baseada em Dy, componente associado ao determinante (Log¦D¦), soma dos dois componentes (Log L+Log¦D¦) a qual propicia a verossimilhança perfilada, e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr), para diferentes valores do coeficiente de competição (β) associados aos dados de cana-de-açúcar. Um modelo com ambos, erros correlacionados espacialmente e efeitos de competição via o modelo fenotípico, foi usado.

Valores de β	**Log L**	**Log¦D¦**	**LogL+Log¦D¦**	**ARc**	**ARr**
-0.80	-1017.61	-31.07	-1048.68	0.43**	0.51**
-0.75	-1017.06	-26.21	-1043.27	0.42**	0.49**
-0.70	-1016.71	-22.05	-1038.76	0.40**	0.48**
-0.65	-1016.56	-18.46	-1035.02	0.38**	0.46**
-0.60	-1016.59	-15.33	-1031.92	0.36**	0.43**
-0.55	-1016.80	-12.60	-1029.40	0.34**	0.41**
-0.50	-1017.15	-10.22	-1027.37	0.31**	0.38**
-0.45	-1017.64	-8.14	-1025.78	0.29**	0.35**
-0.40	-1018.24	-6.34	-1024.58	0.25**	0.32**
-0.35	-1018.91	-4.79	-1023.70	0.22**	0.29**
-0.30	-1019.62	-3.49	-1023.11	0.18**	0.26**
-0.25	-1020.35	-2.40	-1022.75	0.14 ns	0.23**
-0.20	-1021.06	-1.53	-1022.59	0.10 ns	0.19*
-0.15	**-1021.72**	**-0.85**	**-1022.57**	**0.05 ns**	**0.15 ns**
-0.10	-1022.29	-0.38	-1022.67	0.00 ns	0.12 ns
-0.05	-1022.74	-0.09	-1022.83	-0.06 ns	0.07 ns
0.00	-1023.05	0.00	-1023.05	-0.12 ns	0.03 ns
0.05	-1023.21	-0.09	-1023.30	-0.18**	0.00 ns
0.10	-1023.21	-0.38	-1023.59	-0.23**	0.05 ns
0.20	-1022.83	-1.53	-1024.36	-0.33**	-0.14 ns
0.40	-1021.98	-6.34	-1028.32	-0.48**	-0.29**

A Tabela 15 apresenta a verossimilhança perfilada ajustada para uma amplitude de coeficientes de competição (β) em um modelo com ambos, erros correlacionados espacialmente e efeitos de competição. Pode ser visto que a maximização da função de verossimilhança ocorreu para β = -0.15, com um LogL+Log¦D¦ = - 1022.57. Os coeficientes de autocorrelação residual não foram significativos, mostrando que o coeficiente de competição fenotípica englobou todo o padrão de correlação, incluindo os efeitos genéticos de competição e um balanço entre efeitos de competição no nível residual e tendência ambiental dentro de blocos.

A Tabela 16 apresenta a verossimilhança perfilada ajustada para uma amplitude de coeficientes de competição (β) em um modelo somente com os efeitos de competição fenotípica, ou seja, sem erros correlacionados. Pode ser visto que a maximização da função de verossimilhança ocorreu para β = -0.05, com um LogL+Log¦D¦ = - 1023.28. Os dois valores de LogL+Log¦D¦ mencionados são próximos entre si e os resultados confirmam que o modelo somente com os efeitos de competição fenotípica já é suficiente. Pode-se concluir também que os efeitos de competição presentes no ensaio são de pequena magnitude.

Tabela 16. Máximo do logaritmo da função de verossimilhança residual (Log L) baseada em Dy, componente associado ao determinante (Log¦D¦), soma dos dois componentes (Log L + Log¦D¦) a qual propicia a verossimilhança perfilada, e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr), para diferentes valores do coeficiente de competição (β) associados aos dados de cana-de-açúcar. Um modelo somente com efeitos de competição via o modelo fenotípico foi usado.

Valores β	**Log L**	**Log¦D¦**	**LogL+Log¦D¦**
-0.40	-1030.34	-6.34	-1036.68
-0.30	-1026.46	-3.49	-1029.95
-0.20	-1023.95	-1.53	-1025.48
-0.15	**-1023.28**	**-0.85**	**-1024.13**
-0.10	-1023.03	-0.38	-1023.41
-0.05	**-1023.19**	**-0.09**	**-1023.28**
0.00	-1023.77	0.00	-1023.77

A Tabela 17 apresenta resultados comparativos de uma série de modelos aplicados aos dados de cana-de-açúcar.

Tabela 17. Máximo do logaritmo da função de verossimilhança residual (Log L) e estimativas da variância genética entre tratamentos ($\hat{\sigma}^2_\tau$), variância residual ($\hat{\sigma}^2$), herdabilidade ajustada ($\hat{h}^2_{adj}$), coeficientes de competição ($\hat{\beta}$) e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr). Dados de cana-de-açúcar.

Modelo	**Log L**	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2$	$\hat{h}^2_{adj}$	$\hat{\beta}$	**ARc**	**ARr**
Tradicional	-1023.77	685.98	228.73	0.7499	-	-	-
Espacial	-1023.05	684.17	229.26	0.7490	-	-0.12 ^{ns}	0.035^{ns}
Competição (Veros. Perfilada)	-1023.28	667.48	231.43	0.7425	-0.05	-	-
Competição+Espacial (V Perf)	-1022.57	660.02	232.96	0.7391	-0.15	0.05 ^{ns}	0.15 ^{ns}
Competição (Covariável)	-1025.55	674.03	230.56	0.7451	-0.10*	-	-
Competição + Spatial (Covariável)	-1018.97	466.27	347.69	0.5728	-0.64*	0.38*	0.45*

Os modelos tradicional, espacial (auto-regressivo em duas dimensões), de competição fenotípica usando verossimilhança perfilada e de competição via verossimilhança perfilada mais análise espacial deram basicamente o mesmo resultado em termos do Log L, variância residual e herdabilidade. Isto é devido às pequenas magnitudes dos efeitos de competição.

O modelo (3) usando a média dos quatro vizinhos horizontais e verticais como uma covariável tradicional (efeito fixo) e os efeitos de tratamentos como aleatórios confirmaram a presença de efeitos de competição ($\hat{\beta}$ = -0.10). Tal modelo forneceu também a mesma herdabilidade do modelo tradicional e dos modelos espaciais. No entanto, o procedimento REML ordinário aplicado nesse modelo é inadequado porque o coeficiente de competição aparece em ambos, média e variância de y e então não pode ser ajustado como covariável. O procedimento da

verossimilhança perfilada propicia um ajustamento exato e preciso para tal modelo e melhora a estimação dos parâmetros de variância e de competição. As estimativas dos coeficientes de competição mudaram de –0.10 com o REML ordinário para –0.05 com o REML perfilado.

Para o modelo de competição + espacial (última linha da Tabela 17), o procedimento REML ordinário usando uma covariável produziu resultados muito diferentes para Log L, variância residual e herdabilidade. As estimativas dos parâmetros de autocorrelação e do coeficiente de competição foram consideravelmente maiores do que aqueles obtidos com o REML perfilado. Essa diferença revela a importância do uso procedimento REML perfilado, o qual é mais acurado. O modelo de competição + espacial, usando covariável (modelo 11) propiciou um coeficiente de competição igual a -0,64 contra -0,15 propiciado pelo REML perfilado. As estimativas dos parâmetros de autocorrelação foram positivas e altas (0,38 e 0,45), pois nesse caso estão modelando tendência espacial. Esses resultados relatados foram obtidos por meio do uso de valores iniciais positivos para as auto-correlações no processo iterativo. Entretanto, quando foram usados valores iniciais negativos, obtiveram-se valores diferentes de estimativas na convergência. Os valores obtidos foram de 0.40 para $\hat{\beta}$ e -0.47 e -0.29 para as auto-correlações. Tais estimativas não têm sentido porque apresentam sinais opostos aos esperados. Assim, os resultados confirmaram a inconsistência e inadequação da abordagem tradicional de covariável no presente caso.

Os efeitos de genótipos devem ser corrigidos pelo uso da expressão $\tau_c = \tau /(1-\hat{\beta})$. Nesse caso, pelo modelo melhor e parcimonioso, o coeficiente de competição foi –0.05 e então os efeitos de clones devem ser divididos por 1.05 ou multiplicados por 0.952. Isto é equivalente a reduzir a herdabilidade da média de clone, multiplicando-a por 0.952.

4.2 Modelo Genotípico e Fenotípico de Competição em *Eucalyptus maculata*

Modelos de competição genotípica e fenotípica em *E. maculata*, avaliado para o caráter diâmetro da árvore aos 18 anos em experimento no delineamento de blocos ao acaso com 25 famílias e 36 repetições, são apresentados nas Tabelas 18, 19 e 20.

Tabela 18. Máximo do logaritmo da função de verossimilhança residual (Log L) baseada em Dy, componente associado ao determinante (log¦d¦), soma dos dois componentes (Log L + Log¦D¦) a qual propicia a verossimilhança perfilada, e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr), para diferentes valores do coeficiente de competição (β) associados aos dados de *Eucalyptus maculata*. Um modelo com ambos, efeitos espaciais e de competição via o modelo fenotípico, foi usado.

Valores β	Log L	Log¦D¦	LogL+Log¦D¦	ARc	ARr
-0.60	-2924.59	-47.53	-2972.120	0.25^{**}	0.25^{**}
-0.50	-2924.19	-31.71	-2955.900	0.19^{**}	0.20^{**}
-0.40	-2925.21	-19.69	-2944.900	0.13^{**}	0.14^{**}
-0.30	-2927.10	-10.83	-2937.930	0.07^{*}	0.08^{*}
-0.20	-2929.51	-4.742	-2934.250	0.01^{ns}	0.02^{ns}
-0.15	-2930.83	-2.654	-2933.484	-0.02^{ns}	-0.01^{ns}
-0.10	**-2932.20**	**-1.175**	**-2933.375**	$\mathbf{-0.05^{ns}}$	$\mathbf{-0.04^{ns}}$
-0.05	-2933.64	-0.293	-2933.933	-0.07^{*}	-0.07^{*}
0.00	-2935.12	0.000	-2935.120	-0.10^{**}	-0.10^{**}
0.10	-2938.30	-1.175	-2939.480	-0.15^{**}	-0.15^{**}

A Tabela 18 apresenta a verossimilhança perfilada ajustada para uma amplitude de coeficientes de competição (β) em um modelo com ambos, erros espaciais correlacionados e efeitos de competição fenotípica. Pode ser visto que a maximização da função de verossimilhança ocorreu para β = -0.10, com um LogL+Log¦D¦ = -2933.38. Os coeficientes de autocorrelação residual não foram significativos, mostrando que o coeficiente de competição fenotípica englobou todo o padrão de correlação, incluindo os efeitos genéticos de competição e um balanço entre efeitos de competição no nível residual e tendência ambiental dentro de blocos. Esses resultados são coincidentes e análogos àqueles obtidos com a cana-de-açúcar.

A Tabela 19 apresenta a verossimilhança perfilada ajustada para uma amplitude de coeficientes de competição (β) em um modelo somente com efeitos de competição fenotípica, ou seja, sem erros correlacionados. Pode ser visto que a maximização da função de verossimilhança ocorreu para β = -0.10, com um LogL+Log¦D¦ = -2934.30. Os dois valores mencionados de LogL+Log¦D¦ são próximos e os resultados confirmam que o modelo somente com os efeitos de competição fenotípica é suficiente. Enfim, pode-se inferir que os efeitos de competição estão presentes nesse ensaio.

Tabela 19. Máximo do logaritmo da função de verossimilhança residual (Log L) baseada em Dy, componente associado ao determinante (Log¦D¦), soma dos dois componentes (Log L + Log¦D¦) a qual propicia a verossimilhança perfilada, e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr), para diferentes valores do coeficiente de competição (β) associados aos dados de *Eucalyptus maculata*. Um modelo somente com efeitos de competição via o modelo fenotípico foi usado.

Valores β	**Log L**	**Log¦D¦**	**LogL+Log¦D¦**
-0.40	-2933.55	-19.69	-2953.240
-0.30	-2929.77	-10.83	-2940.600
-0.20	-2929.63	-4.742	-2934.370
-0.10	**-2933.13**	**-1.175**	**-2934.305**
-0.05	-2936.22	-0.293	-2936.513
0.00	-2940.18	0.000	-2940.180
0.10	-2950.59	-1.175	-2951.770

A Tabela 20 apresenta resultados comparativos de uma série de modelos aplicados aos dados de *E. maculata.*

Tabela 20. Máximo do logaritmo da função de verossimilhança residual (Log L) e estimativas da variância genética entre tratamentos ($\hat{\sigma}^2_\tau$), variância residual ($\hat{\sigma}^2$), herdabilidade ajustada ($\hat{h}^2_{adj}$), coeficientes de competição ($\hat{\beta}$) e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr). Dados de *Eucalyptus maculata.*

Modelo	Log L	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2$	$\hat{h}^2_{adj}$	$\hat{\beta}$	ARc	ARr
(a) Tradicional	-2940.18	37.25	601.44	0.233	-	-	-
(b) Espacial	-2935.12	35.80	596.61	0.226	-	-0.10**	-0.10**
(c) Competição (Veros. Perfilada)	-2934.30	36.28	590.72	0.231	-0.10	-	-
(d) Competição + Esp. (Ver. Perf.)	-2933.38	35.83	588.00	0.229	-0.10	-0.05^{ns}	-0.04 ^{ns}
(e) Competição (Covariável)	-2932.19	34.82	585.97	0.224	-0.25**	-	-
(f) Espacial+ Competição (Cov.)	-2927.19	30.34	644.52	0.180	-0.52**	0.21**	0.21**
(g) Espacial+ Competição Genotíp.	-2935.12	35.80	596.61	0.226		-0.10**	-0.10**

O modelo espacial mostrou-se melhor que o tradicional em termos do Log L. Esses, mais os modelos de competição (usando verossimilhança perfilada) e de competição (usando verossimilhança perfilada) + espacial produziram basicamente os mesmos resultados em termos de variância residual e herdabilidade. O ajuste para os efeitos de competição não reduziu a estimativa da herdabilidade. Isto ocorreu porque nesse conjunto de dados a competição manifestou-se apenas no nível residual (discutido posteriormente), ou seja, aqui a competição é meramente um efeito ambiental.

O modelo (3) usando a média dos quatro vizinhos horizontais e verticais como uma covariável tradicional (efeito fixo) e os efeitos de tratamentos como aleatórios confirmaram a presença de efeitos de competição ($\hat{\beta}$ = -0.25). Propiciou também a mesma herdabilidade que os modelos tradicional e espacial. No entanto, o procedimento REML ordinário aplicado nesse modelo é inadequado, conforme comentado anteriormente. As estimativas dos coeficientes de competição mudaram de –0.25 com o REML ordinário para –0.10 com o REML perfilado, confirmando a superestimação dos efeitos de competição.

Para o modelo de competição + espacial, o procedimento REML ordinário usando uma covariável produziu resultados muito diferentes para Log L, variância residual e herdabilidade. As estimativas dos parâmetros de autocorrelação e do coeficiente de competição foram consideravelmente maiores do que aqueles obtidos com o REML perfilado. O modelo de competição + espacial, usando covariável (modelo f na Tabela 20), propiciou um coeficiente de competição igual a -0,52 contra -0,10 propiciado pelo REML perfilado. As estimativas dos parâmetros de autocorrelação foram positivas e moderadas (0,21 e 0,21), pois nesse caso estão modelando tendência espacial.

O modelo de competição genotípica + espacial forneceu os mesmos resultados que o modelo espacial, revelando nenhuma significância dos efeitos genéticos de competição (Tabela 20). Assim, uma estimativa plausível do coeficiente de competição é –0,10 (obtido via verossimilhança perfilada) e os efeitos de competição são plenamente considerados no modelo espacial. Pode ser observado também que as estimativas dos parâmetros de autocorrelação pelo modelo espacial foram também iguais a –0,10. Quando aplicados separadamente sobre vizinhos nas linhas e nas colunas, as estimativas de ambos coeficientes de competição foram também iguais a –0,10. Isto mostra que, no caso de ausência de competição no nível genético, os modelos espaciais e de competição fenotípica (modelo c na Tabela 20) modelam os mesmos efeitos, os quais são equivalentes a um balanço entre competição no nível residual e tendência ambiental residual.

Tendência ambiental e competição residual são efeitos confundidos e não podem ser separados. No entanto, não existe uma razão prática para essa separação. Os modelos espacial e de competição fenotípica diferem somente na presença de competição em nível genético (caso do conjunto de dados de pínus, discutido posteriormente).

Uma comparação envolvendo os modelos tradicional, espacial e de competição fenotípica, em termos do ordenamento de genótipos, é apresentado na Tabela 21. Pode ser visto que os três modelos produziram ordenamentos e efeitos genotípicos preditos muito similares. Os mesmos genótipos podem ser selecionados pelos três modelos usando as intensidades de seleção de 20 % (seleção dos 5 melhores) ou 50 % (seleção dos 13 melhores). Esse resultado é esperado no caso de baixos coeficientes de competição como aqueles obtidos na presente aplicação (-0.10). Usando simulação, Kusnandar (2001) relatou que os modelos incluindo competição não conduziram a

melhores resultados quando as magnitudes dos parâmetros de competição foram baixas (entre 0 e 0,10, em valor absoluto). De acordo com o autor, os modelos incluindo competição tornaram-se vantajosos quando os coeficientes de competição foram menores que -0,30.

Os efeitos genotípicos de tratamentos devem ser corrigidos pela expressão $\tau_c = \tau /(1-\hat{\beta})$. Nesse caso, o coeficiente de competição foi –0,10 e então os efeitos genotípicos pelo modelo de competição (e também pelo modelo espacial) na Tabela 21 devem ser divididos por 1,10 ou multiplicados por 0,91. Isto é equivalente a multiplicar a herdabilidade da média de tratamentos por 0,91. Por ambos, análise tradicional e modelo de competição, tal herdabilidade equivaleu a 0,69. Multiplicando esse valor por 0,91 produz-se 0,63, o qual é menor e mais realístico do que o 0,69 obtido com a análise tradicional. Então, o uso dos modelos de competição e espacial nesse caso tem maior importância na estimação de ganhos genéticos do que na propriedade de ordenamento dos materiais genéticos para a seleção.

Tabela 21. Comparação envolvendo os modelos tradicional, espacial e de competição em termos do ordenamento de genótipos e seus efeitos preditos. Dados de *Eucalyptus maculata.*

	Ordenamento de Famílias			**Efeitos Preditos de Famílias**		
Famílias	**Competição**	**Espacial**	**Tradicional**	**Competição**	**Espacial**	**Tradicional**
579	1	1	1	10.26	10.12	10.78
565	2	2	2	8.104	8.281	8.153
572	3	3	3	6.854	6.993	6.933
580	4	5	4	5.042	4.522	5.263
577	5	4	5	4.786	4.954	4.653
573	6	7	6	2.384	1.509	2.558
576	7	12	7	2.136	2.453	2.194
584	8	8	10	1.556	1.858	1.438
561	9	9	9	1.552	1.620	1.490
562	10	6	8	1.334	1.276	1.553
581	11	10	12	1.184	0.975	1.110
563	12	11	11	1.166	2.407	1.122
574	13	13	13	0.629	-0.122	0.680

Os modelos de competição usando verossimilhança perfilada ajustada em ambos os conjuntos de dados, cana-de-açúcar e Eucalyptus, produziram resultados coerentes em termos da não significância das estimativas dos parâmetros autoregressivos no modelo simultâneo de competição fenotípica e análise espacial. Esse resultado era esperado, uma vez que o ajuste para os efeitos de competição fenotípica contempla basicamente as mesmas fontes de variação que os parâmetros autoregressivos, quando não existe competição em nível genético. Nessa situação, os modelos de competição fenotípica e os modelos espaciais provavelmente fornecem os mesmos resultados. Em ausência de competição no nível genético, o modelo de competição fenotípica torna-se o próprio método de Papadakis e é esperado produzir os mesmos resultados que as abordagens de Papadakis (1937), Bartelett (1978) e Kempton e Howes (1981) usadas para controle da tendência espacial em fertilidade. Uma vez que o modelo autoregressivo separável em duas dimensões engloba o método de Papadakis (Gilmour et al., 1997), espera-se que o modelo de competição fenotípica e o modelo espacial conduzam aos mesmos resultados em ausência de competição genotípica. Tais resultados não foram atingidos pelo procedimento REML ordinário. É importante relatar que o uso da verossimilhança perfilada associada ao modelo de competição fenotípica é um procedimento melhorado em relação ao método de Papadakis. No ajuste do método de Papadakis e do modelo autoregressivo separável em duas dimensões, uma mistura de competição em nível residual e tendência ambiental local está sendo modelada. Correll e Anderson (1983) relataram que os termos de Papadakis e de competição intergenotípica foram efetivamente não correlacionadas. Isto é esperado, pois os componentes residual e genético da competição são termos independentes e, portanto, não correlacionados.

Em termos paramétricos, o efeito de competição de uma planta i é dado por $c_i = \phi_i + \gamma_i$, em que ϕ_i é o efeito de competição genotípica e γ_i é o efeito de competição residual. O modelo paramétrico para o efeito residual total é dado por $e_i = \gamma_i + \xi_i + \eta_i$ e então o modelo paramétrico para o fenótipo total (em termos de um vetor) pode ser decomposto em $y = Xb + Z\tau + NZ\phi + \gamma + \xi + \eta$. O modelo de competição fenotípica trata os elementos ϕ_i, γ_i, ξ_i e η_i todos em conjunto em um único componente $\beta = \phi_i + \gamma_i + \xi_i + \eta_i$. O modelo espacial autoregressivo considera $e_i = \gamma_i + \xi_i + \eta_i$. Dessas fórmulas pode ser visto que os modelos espacial autoregressivo e de competição fenotípica são

idênticos em ausência de competição no nível genético. Em geral, os seguintes modelos são ótimos (em termos da consideração de todos os efeitos especificados no modelo para o fenótipo total) nas seguintes situações:

(i) modelo espacial autoregressivo: ótimo em ausência de competição no nível genético;

(ii) modelo fenotípico de competição via verossimilhança perfilada: ótimo em qualquer situação;

(iii) modelo fenotípico de competição via verossimilhança perfilada mais modelo espacial autoregressivo: ótimo em qualquer situação, uma vez que tende a ser igual a (ii). Porém tende a ser superparametrizado;

(iv) modelo genotípico de competição mais modelo espacial autoregressivo: ótimo em qualquer situação;

(v) modelo genotípico de competição: ótimo em ausência de competição residual e tendência ambiental espacial.

Pode ser dito também que os modelos de competição somente são necessários quando a competição tem uma base genética. Caso contrário, os modelos tradicional e/ou espacial autoregressivo são suficientes. Então, recomenda-se verificar a significância dos efeitos genotípicos de competição como um primeiro passo na análise. Esse resultado guiará o estatístico à escolha de melhores modelos para análises subseqüentes.

Na presença de efeitos genéticos de competição, existem duas opções: (a) uso do modelo genotípico de competição mais modelo espacial autoregressivo; (b) uso do modelo fenotípico de competição via verossimilhança perfilada. Esse último modelo considera implicitamente três efeitos: competição genotípica, competição residual e tendência ambiental. O modelo em (a) considera explicitamente a competição genotípica e também a covariância entre efeitos genéticos diretos e efeitos de competição sobre os vizinhos. Então, tal modelo é mais preciso e deve ser usado para aplicações práticas.

4.3 Modelos Genotípico e Fenotípico de Competição em *Pinus*

Modelos de competição genotípica e fenotípica em *Pinus*, avaliado para o caráter diâmetro aos 13 anos em experimento no delineamento em látice com 121 famílias e 6 repetições com 6 plantas por parcela, são apresentados nas Tabela 22.

Tabela 22. Máximo do logaritmo da função de verossimilhança residual (Log L) e estimativas da variância genética entre tratamentos ($\hat{\sigma}^2_\tau$), variância residual ($\hat{\sigma}^2$), herdabilidade ajustada ($\hat{h}^2_{adj}$), coeficientes de competição ($\hat{\beta}$) e coeficientes de autocorrelação associados a colunas (ARc) e linhas (ARr). Dados de Pínus.

Modelo	Log L	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2$	$\hat{h}^2_{adj}$	$\hat{\beta}$	ARc	ARr
(a) Tradicional	-6584.67	1.0403	18.255	0.2156	-	-	-
(b) Espacial	-6559.19	0.9621	17.891	0.2040	-	-0.10*	-0.13*
(c)Competição+Espacial (Ver. Perf.)	-6512.25	1.1174	16.975	0.2470	-0.18*	-0.03 ns	-0.05 ns
(d) Competição (Covariável)	-6498.27	1.1795	16.960	0.2600	-0.23*	-	-
(e) Competição+Espacial(Cov)	-6496.79	1.1497	16.945	0.2541	-0.22*	-0.01ns	-0.04ns
(f) Competição Gen.-Norte	-6574.99	1.0515	18.192	0.2186	-	-	-
(g) Espacial + Gen. Norte	-6547.81	0.9837	17.831	0.2091	-	-0.10*	-0.13*
(h) Espacial + Gen. Norte + Cov Gen.	-6543.87	0.9476	17.779	0.2024	-	-0.10*	-0.13*

O modelo espacial apresentou melhor ajuste que o tradicional e revelou a presença de competição, conforme significância dos negativos coeficientes de autocorrelação associados a linhas e colunas. A presença de competição foi confirmada pela significância do coeficiente de competição em (c) na Tabela 22, em um modelo que inclui também erros espaciais. Esse modelo, ajustado via verossimilhança perfilada, não apresentou significância dos coeficientes de autocorrelação. Os coeficientes de competição foram superestimados pela abordagem da covariável (modelos d e e na Tabela 22). Pode ser visto que o modelo de competição fenotípica difere (parâmetro autoregressivo

de competição maior do que os parâmetros espaciais autoregressivos) do modelo espacial somente em presença de competição em nível genético. Isto ocorreu no presente conjunto de dados, mas não nos exemplos prévios.

A consideração da competição em ambos os níveis, genotípico e residual, pode ser vantajosa. Isto foi realizado pelos modelos (g) e (h). Inicialmente, um modelo sem termos espaciais, mas incluindo individualmente oito competidores, foi avaliado. Esse modelo revelou a significância apenas dos efeitos dos vizinhos na posição norte em termos da competição genotípica. Assim, um modelo incluindo apenas essa posição foi ajustado em (f) na Tabela 22. Esse modelo demonstrou ser intermediário entre os modelos tradicional e espacial em (a) e (b), respectivamente, conforme pode ser visto pelos Log L. Então, a competição residual mostrou-se maior do que a competição genotípica.

A modelagem simultânea da competição genotípica (posição norte) e residual de acordo com o modelo (g) apresentou melhor ajuste e mostrou os mesmos valores de coeficientes de autocorrelação associados a linhas e colunas que o modelo espacial em (b). Isto confirma que a análise espacial estava modelando competição apenas no nível residual e que isto não é suficiente nesse caso, em que a competição é também devida a causas genéticas.

Um modelo mais completo, permitindo também a covariância entre efeitos genéticos diretos e efeitos de competição sobre os vizinhos, foi ajustado em (h) na Tabela 22. Esse modelo apresentou melhor ajuste que o modelo (g), sem tal covariância. Também propiciou uma menor estimativa de herdabilidade, conforme esperado sob ajuste para competição e, provou estar modelando adequadamente a competição em ambos os níveis. O mesmo modelo revelou uma correlação genética negativa (-0,68) entre efeitos diretos e sobre os vizinhos. Isto revela a mesma tendência verificada com base no coeficiente de competição fenotípica. O modelo forneceu ainda uma herdabilidade de 7,% para os efeitos indiretos sobre os vizinhos, isto é, herdabilidade para os efeitos de competição. Os efeitos significativos apenas na posição norte são provavelmente associados ao sombreamento ditado pelo caminhamento do sol na região.

Uma comparação explícita formal entre os modelos fenotípico espacial (c) e genotípico espacial (h) não pode ser realizada uma vez que eles contêm diferentes efeitos fixos. Teoricamente e conceitualmente, o modelo genotípico é mais completo. Esses modelos foram comparados em

termos de ordenamento de genótipos e ganho genético. Tomando o modelo (h) como o melhor e correto, foram verificadas as seguintes taxas de coincidências com os demais modelos, para a intensidade de seleção de 10 % das melhores famílias: 91,7 % com o modelo (c), 83,3 % com o modelo (b) e 75 % com o modelo (a). Assim, a eficiência seletiva do modelo de competição fenotípica foi próxima daquela obtida com o modelo de competição genotípica. No entanto, os ganhos genéticos estimados foram de 5,68 % pelo modelo fenotípico e 4,37 % pelo modelo genotípico, significando uma superestimação de 30 % pelo modelo (c), conforme esperado com base na herdabilidade estimada por tal modelo.

A Tabela 23 mostra correlações genéticas negativas entre os efeitos genéticos diretos no próprio indivíduo e indiretos nos vizinhos obtidos por meio do modelo (g). Os altos e negativos efeitos de vizinhança associados às melhores famílias mostram que as mesmas são muito agressivas e tiveram seus valores reais superestimados nos modelos sem os efeitos de competição genotípica. Isto mostra a ineficiência de modelos espaciais e não espaciais simples quando existe competição genotípica.

Tabela 23. Efeitos genotípicos preditos para as melhores famílias de Pínus, por meio do modelo genotípico de competição mais modelo espacial autoregressivo.

Ordem pelos Efeitos Totais	Genótipos	Ordenamento de Genótipos		
		Efeitos Diretos (τ_i)	Efeitos Indiretos (ϕ_i)	Efeitos Totais ($\tau_i + \phi_i$)
1	98	3.251	-1.238	2.013
2	96	2.034	-0.646	1.388
3	70	2.018	-0.932	1.086
4	20	1.061	-0.203	0.858
5	25	1.447	-0.745	0.702
6	66	0.839	-0.158	0.681
7	106	0.842	-0.186	0.656
8	99	1.046	-0.406	0.640
9	69	0.735	-0.176	0.558
10	21	0.504	0.044	0.547
11	45	0.612	-0.060	0.546
12	107	1.044	-0.499	0.544

A seleção deve então ser baseada em $\hat{\tau} + \hat{\phi}$, em que $\hat{\phi}$ é negativo nas variedades mais agressivas. Verifica-se que houve inversão no ordenamento quando se compara a seleção baseada nos efeitos diretos e nos efeitos totais. Por exemplo, o genótipo 20 muda da quinta para a quarta posição. A seleção pode basear-se também em $\hat{\phi}$, visando à identificação de genótipos adequados a plantios adensados como, por exemplo, no melhoramento do cafeeiro e do dendezeiro. Nesse caso, o melhor genótipo foi o 20, pois dentre os melhores, é o que apresenta menor efeito adverso sobre os seus vizinhos, ou seja, é o que apresentará menor depressão na produtividade, quando cultivado em plantios puros.

CAPÍTULO 7
ANÁLISE MULTIVARIADA, DIVERGÊNCIA GENÉTICA E ÍNDICES DE SELEÇÃO

1 TÉCNICAS DE ANÁLISE MULTIVARIADA

As técnicas de Análise Multivariada são muito úteis na análise experimental, principalmente nas áreas de biometria (incluindo a agrometria), psicometria, quimiometria e tecnometria. Tais técnicas multivariadas podem ser divididas em dois grupos, segundo Kendall (1950):

(i) Técnicas de análise da interdependência e relações entre si de um conjunto de variáveis: análise de agrupamento, análise de componentes principais e análise de fatores.

(ii) Técnicas para análise da dependência de uma ou mais variáveis em função das outras: análise discriminante, análise de variância multivariada, análise de correlações canônicas, análise de regressão multivariada, análise de medidas repetidas sob modelo de efeitos fixos.

Por outro lado, Johnson e Wichern (1988) resumem as principais técnicas de análise multivariada da forma apresentada na seqüência.

1. Métodos para Distinção entre Grupos

   a) Análise de Agrupamento

   b) Análise Discriminante

2. Métodos para Estudo da Estrutura de Covariância ou Correlação entre Variáveis

   c) Componentes Principais

   d) Análise de Fatores

Dentre estas quatro técnicas gerais, três (a, c e d) se relacionam mais diretamente com as aplicações em genética e melhoramento. A análise discriminante (b) tem como maior aplicação a discriminação ou alocação de um conjunto de genótipos em grupos ou populações previamente conhecidos, usando para isto um certo número de características avaliadas. Assim, as técnicas pertencentes ao grupo (i) na classificação de Kendall e aos tópicos (a), (c) e (d) na classificação de Johnson e Wichern, serão prioritariamente abordados aqui. Tais técnicas serão inicialmente abordadas no contexto tradicional e em seguida em conexão com a metodologia de modelos mistos. Adicionalmente, será considerada também a própria análise multivariada não estruturada de modelos mistos associada à técnica dos índices de seleção e também a técnica da transformação canônica de variáveis visando substituir tal análise multivariada não estruturada por várias análises univariadas. Também, a transformação Cholesky é abordada visando à obtenção de variáveis não correlacionadas residualmente e com variância residual unitária.

A análise de agrupamento permite a formação de grupos (não conhecidos previamente) por meio de técnicas de agrupamento aplicadas sobre medidas de dissimilaridade entre fenótipos. Várias medidas podem ser usadas, destacando-se as distâncias fenotípicas tais quais a Euclidiana (com algumas variações ou tipos) e a distância estatística ou de Mahalanobis (Cruz e Regazzi, 1994; Cruz e Carneiro, 2003). Sob modelos com efeitos aleatórios de tratamentos, os valores genéticos preditos devem ser usados em lugar dos valores fenotípicos. Também uma matriz de variâncias e covariâncias dos valores genéticos deve ser usada para cômputo da distância de Mahalanobis, em

lugar da matriz de dispersão residual usada nos modelos com efeitos fixos de tratamento. É também desejável que os valores genéticos para todos os caracteres sejam preditos simultaneamente por um modelo misto multivariado. Nesse último caso, não se justifica usar a distância de Mahalanobis já que todas as correlações já terão sido usadas no processo de predição multivariada.

A análise de componentes principais (PCA) e a análise de fatores (FA) permitem simplificar a estrutura multivariada (n caracteres) dos dados e, posteriormente, permitem a dispersão gráfica dos indivíduos ou genótipos em dois ou três eixos coordenados e, portanto, permitem a visualização de grupos e genótipos mais e menos divergentes. No caso, a maioria da variação no espaço n dimensional é explicada no espaço bi ou tri-dimensional (dois ou três eixos). Em outras palavras, as n variáveis originais são substituídas por dois ou três componentes principais ou fatores, dependendo da técnica empregada (PCA ou FA). Conforme Cruz e Carneiro (2003), uma variação da técnica de componentes principais é a análise de variáveis canônicas, a qual é aplicada quando se dispõe de informações dentro de acessos, ou seja, repetições experimentais. Neste caso, usa-se uma matriz de dispersão residual à semelhança do que é realizado no cômputo da distância de Mahalanobis. Quando aplicada sobre valores genotípicos preditos, a PCA não necessita considerar a matriz de dispersão residual pois a mesma já terá sido considerada na ocasião da predição dos valores genotípicos. Assim, a PCA poderá ser usada de maneira eficiente, considerando apenas as matrizes de valores genéticos preditos padronizados e de correlações genéticas entre os caracteres.

A análise de fatores pode ser considerada como uma extensão da análise de componentes principais. Na PCA o vetor de valores fenotípicos é dado por y = u + g + e, em que u é a média geral, g é o vetor de efeitos genotípicos e e é o vetor de erros aleatórios. Na AF, g é desdobrado em $g = \Lambda f + \delta$, em que f denota o vetor dos escores fatoriais, Λ é a matriz dos carregamentos nos fatores ou cargas fatoriais e δ é um vetor aleatório de erros específicos representando a falta de ajustamento do modelo fatorial. Então, na FA, $y = u + \Lambda f + \delta + e$. Portanto, a FA baseia-se em um modelo estatístico propriamente dito para os efeitos genotípicos (g), o que é uma vantagem sobre a PCA. Neste modelo estatístico, suposições de normalidade são feitas para os efeitos aleatórios f e δ.

Dado o modelo aleatório associado à técnica AF, a mesma pode ser adotada no contexto dos modelos mistos com genótipos aleatórios por meio dos modelos mistos fator analíticos (FAMM). Os modelos FAMM são uma regressão aleatória multivariada, com coeficientes de regressão e covariáveis desconhecidos, e, portanto, ambos devem ser estimados. Tais modelos podem ser aplicados para o caso de múltiplos caracteres e também múltiplos experimentos no contexto da interação genótipos x ambientes. Para o caso de medidas repetidas, a técnica da regressão aleatória com covariáveis conhecidas pode ser aplicada para simplificação da estrutura multivariada. No caso, as covariáveis são os tempos ou idades em que as observações são tomadas, ao passo que os coeficientes de regressão devem ser estimados. Outras opções para o caso de medidas repetidas são os modelos SAD e ARH (ver Capítulo 9).

Para o caso de múltiplos caracteres, o modelo FAMM deve ser aplicado aos efeitos de genótipos, e estruturas multivariadas devem ser aplicadas aos demais efeitos aleatórios (parcela e erro). Isto difere do FAMM para interação genótipo x ambiente, em que os demais efeitos aleatórios do modelo são não correlacionados entre locais e, portanto, não demandam estrutura multivariada. Os escores fatoriais (BLUP's) dos genótipos podem ser plotados para os fatores 1 e 2 (ou 1, 2 e 3), permitindo agrupar genótipos com base na similaridade ou, em outras palavras, separar genótipos com base na divergência. Também, a técnica do biplot pode ser usada em associação com os modelos FAMM.

A técnica PCA pode também ser aplicada no contexto dos modelos mistos e usada na análise de múltiplos caracteres. Nesse contexto, é denominado componentes principais genéticos ou PCA sob modelos mistos (*PCAM models – principal component analysis under mixed models*). As analogias entre os modelos FAMM e PCAM são discutidas em tópicos seguintes.

2 MODELO MISTO MULTIVARIADO E ÍNDICE DE SELEÇÃO

A análise simultânea de vários caracteres, visando estimar a estrutura de covariância ou correlação e também a predição de valores genéticos para fins de seleção (para caracteres individuais e também por índices de seleção), é realizada de maneira eficiente pelo procedimento

REML/BLUP multivariado ou análise multivariada não estruturada. Neste caso, o modelo multivariado é especificado de forma a contemplar a covariância ambiental existente entre os caracteres (Henderson, 1984; Resende, 2002; Mrode 2005). Assim, este modelo difere do modelo multivariado referido no Capítulo 8. Com valores genéticos preditos por um modelo multivariado, os índices de seleção são estabelecidos via ponderação desses valores genéticos pelos pesos econômicos dos caracteres. As herdabilidades e correlações já terão sido consideradas na predição pelo modelo multivariado. Índices aditivos, multiplicativos e de rank dos valores genéticos preditos podem ser utilizados, conforme implementado no *software* Selegen-Reml/Blup.

Para estimação das correlações genéticas, modelos bivariados tomando as variáveis duas a duas podem ser utilizados, visando evitar problemas de convergência em modelos multivariados com grande número de caracteres.

2.1 Modelo Misto Multivariado

Os modelos multivariados destinam-se à avaliação de indivíduos, simultaneamente, para dois ou mais caracteres e apresentam grande relevância no contexto da seleção envolvendo agregados genotípicos como objetivo da seleção. Tais modelos exploram as correlações genéticas e fenotípicas entre caracteres e foram aplicados inicialmente por Henderson e Quaas (1976).

Esses modelos podem também ser aplicados vantajosamente quando o objetivo da seleção é o melhoramento de um único caráter, nas seguintes situações: (*i*) quando o caráter é avaliado mais de uma vez no indivíduo, mas a correlação genética entre as medições não equivale a 1 (neste caso o modelo multivariado pode ser superior ao modelo de repetibilidade, principalmente quando a correlação for menor que 0,80); (*ii*) quando o caráter é avaliado em parentes em diferentes ambientes e a correlação genética entre a performance nos dois ambientes não é 1; (*iii*) no contexto da seleção empregando caracteres auxiliares no melhoramento, onde se explora a correlação genética e fenotípica, bem como a diferença entre as herdabilidades dos caracteres.

Uma das principais vantagens da utilização dos modelos multivariados consiste no aumento da acurácia das avaliações. Esse aumento pode ocorrer apenas em determinadas situações. Na seleção usando caracteres auxiliares, a acurácia é dada por h_m, a raiz quadrada da herdabilidade multivariada. A herdabilidade multivariada é dada por:

$$h_m^2 = h_y^2 \left[1 + \frac{(h_x \; r_{axy} / h_y - r_{xy})^2}{1 - r_{xy}^2}\right], \; onde \; h_x^2 \; e \; h_y^2$$ referem-se às herdabilidades do caráter auxiliar (x) e do caráter objetivo (y), respectivamente, e r_{axy} e r_{xy} são relativas às correlações genéticas e fenotípicas entre os dois caracteres, respectivamente. Neste caso, a eficiência ou ganho em acurácia com a seleção multivariada é dada pela razão entre as acurácias multivariada e univariada, ou seja:

$$Ef = \frac{h_m}{h_y} = \left[1 + \frac{(h_x \; r_{axy} / h_y - r_{xy})^2}{1 - r_{xy}^2}\right]^{1/2}$$

Por esta expressão, verifica-se:

(i) Se $h_x/h_y \; r_{axy} = r_{xy}$, a eficiência será 1 e a análise multivariada será equivalente à univariada.

(ii) Sendo as herdabilidades iguais ($h_x = h_y$), a eficiência depende apenas da diferença entre as correlações genética e fenotípica. Sendo a correlação fenotípica igual a:

$$r_{xy} = r_{xy} \; h_x \; h_y + r_{Exy} \; [(1 - h_x^2)(1 - h_y^2)]^{1/2}$$

onde r_{Exy} é a correlação ambiental entre os dois caracteres, tem-se que: $(r_{axy} - r_{xy}) = r_{axy} - \{r_{axy} \; h_x \; h_y + r_{Exy} \; [(1 - h_x^2)(1 - h_y^2)]^{1/2}\}$. Por esta última expressão, verifica-se que a maximização (que significa maximizar a eficiência da seleção multivariada) da diferença entre r_{axy} e r_{xy}, só pode se dar por meio da maximização da diferença entre r_{Exy} e r_{xy}. Assim, o ganho em acurácia depende da diferença absoluta entre a correlação genética e a correlação ambiental entre os caracteres.

(iii) Sendo as herdabilidades iguais ($h_x = h_y$) e as correlações genética e ambiental também iguais ($r_{axy} = r_{Exy}$), a análise multivariada será equivalente à análise univariada.

(iv) Para os casos em que a correlação genética for superior à correlação fenotípica, o ganho em acurácia será tanto maior quanto maior a herdabilidade do caráter auxiliar em relação à herdabilidade do caráter objetivo do melhoramento.

Uma desvantagem da análise multivariada ou multicaracterística refere-se ao alto custo computacional. O custo da análise multivariada de n caracteres é maior que o custo de n análises univariadas. Adicionalmente, a análise sob modelo multivariado requer estimativas fidedignas de correlações genéticas e fenotípicas entre os caracteres (Mrode, 1996).

No contexto dos modelos lineares mistos, o seguinte modelo multivariado para a predição dos efeitos aditivos e de dominância pode ser enunciado:

$y = Xb + Za + Zd + Wc + e$, em que:

y, b, a, d, c e *e* : vetores de dados, de efeitos fixos, de efeitos genéticos aditivos (aleatório), de efeitos de dominância (aleatório), de efeitos de parcela (aleatório) e de erros aleatórios, respectivamente.

X, Z, W : matrizes de incidência para *b, a (d)* e *c*, respectivamente.

Considerando dois caracteres (altura e diâmetro), os modelos para cada caráter são:

$y_1 = X_1b_1 + Z_1a_1 + Z_1d_1 + W_1c_1 + e_1$ para altura e

$y_2 = X_2b_2 + Z_2a_2 + Z_2d_2 + W_2c_2 + e_2$ para diâmetro

Estes modelos expressos em notação matricial equivalem a:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} + \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

Para o modelo bivariado, é a seguinte a estrutura de variâncias:

$$Var\begin{bmatrix} a_1 \\ a_2 \\ d_1 \\ d_2 \\ c_1 \\ c_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1}A & \sigma_{a_{12}}A & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{a_{12}}A & \sigma^2_{a_2}A & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{d_1}D & \sigma_{d_{12}}D & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{d_{12}}D & \sigma^2_{d_2}D & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{e_1}I & \sigma_{e_{12}}I & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{e_{12}}I & \sigma^2_{e_2}I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{ad_1}I & \sigma_{ad_{12}}I \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{ad_{12}}I & \sigma^2_{ad_2}I \end{bmatrix}$$, em que:

$\sigma^2_{a_i}$: variância genética aditiva para o caráter *i*.

$\sigma^2_{d_i}$: variância genética de dominância para o caráter *i*.

$\sigma^2_{e_i}$: variância ambiental entre parcelas para o caráter *i*.

σ^2_{ad} : variância ambiental dentro de parcela para o caráter *i*.

$\sigma_{a_{12}}$: covariância genética aditiva entre os caracteres 1 e 2.

$\sigma_{d_{12}}$: covariância genética de dominância entre os caracteres 1 e 2.

$\sigma_{e_{12}}$: covariância ambiental ao nível de parcela entre os caracteres 1 e 2.

$\sigma_{ad_{12}}$: covariância ambiental dentro de parcela entre os caracteres 1 e 2.

A : matriz de parentesco genético aditivo.

D : matriz de parentesco genético de dominância.

As equações de modelo misto para a predição de efeitos aditivos e de dominância pelo procedimento BLUP sob modelo individual correspondem a:

$$\begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{d} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}Z & X'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z+G^{-1} & Z'R^{-1}Z & Z'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z & Z'R^{-1}Z+G_d^{-1} & Z'R^{-1}W \\ W'R^{-1}X & W'R^{-1}Z & W'R^{-1}Z & W'R^{-1}W+C^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ Z'R^{-1}y \\ W'R^{-1}y \end{bmatrix}$$ (1), em que:

$$\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix}; \qquad \hat{a} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \end{bmatrix}; \qquad \hat{d} = \begin{bmatrix} \hat{d}_1 \\ \hat{d}_2 \end{bmatrix}; \qquad \hat{c} = \begin{bmatrix} \hat{c}_1 \\ \hat{c}_2 \end{bmatrix}$$

$$R^{-1} = R_o^{-1} \otimes I; \quad G^{-1} = G_o^{-1} \otimes A^{-1}; \quad G_d^{-1} = G_{do}^{-1} \otimes D^{-1}; \quad C^{-1} = C_o^{-1} \otimes I$$

$$R_o = \begin{bmatrix} \sigma^2_{ad_1} & \sigma_{ad_{12}} \\ \sigma_{ad_{12}} & \sigma^2_{ad_2} \end{bmatrix}; \quad G_o = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix}; \quad G_{do} = \begin{bmatrix} \sigma^2_{d_1} & \sigma_{d_{12}} \\ \sigma_{d_{12}} & \sigma^2_{d_2} \end{bmatrix}; \quad c_o = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

$\otimes$: denota a operação produto de Kronecker.

Se $\sigma_{ad_{12}} = \sigma_{a_{12}} = \sigma_{d_{12}} = \sigma_{e_{12}} = 0$, as equações reduzem-se a análises univariadas dos dois caracteres, considerados como não correlacionados. Por outro lado, se $\sigma_{ad_{12}} = \sigma_{e_{12}} = 0$, as equações tornam-se equivalentes àquelas descritas no Capítulo 8, para a seleção de um mesmo caráter em diferentes ambientes.

Para o caso balanceado, as equações de modelo misto apresentadas, no seu componente aditivo, estão associadas ao seguinte índice multiefeitos para o caráter 1 (Resende et al., 1994b):

$$I_1 = b_1(Y_{ijK} - \overline{Y}_{ij.}) + b_2(X_{ijk} - \overline{X}_{ij.}) + b_3(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}) + b_4(\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...}) +$$

$$+ b_5(\overline{Y}_{i..} - \overline{Y}_{...}) + b_6(\overline{X}_{i..} - \overline{X}_{...}),$$ para o caráter *y*, em que:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \frac{n-1}{n}\ \sigma^2_{\delta_1} & \frac{n-1}{n}\ \sigma_{\delta_{12}} \\ \frac{n-1}{n}\ \sigma_{\delta_{21}} & \frac{n-1}{n}\ \sigma^2_{\delta_2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{n-1}{n}\ (1-\rho_a)\ \sigma^2_{a_1} \\ \frac{n-1}{n}\ (1-\rho_a)\sigma_{a_{12}} \end{bmatrix}$$

$$\begin{bmatrix} b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} \frac{p-1}{p}\ \frac{b-1}{b}(\sigma^2_{e_1} + \frac{\sigma^2_{\delta_1}}{n}) & \frac{p-1}{p}\ \frac{b-1}{b}(\sigma_{e_{12}} + \frac{\sigma_{\delta_{12}}}{n}) \\ \frac{p-1}{p}\ \frac{b-1}{b}(\sigma_{e_{12}} + \frac{\sigma_{\delta_{12}}}{n}) & \frac{p-1}{p}\ \frac{b-1}{b}(\sigma^2_{e_2} + \frac{\sigma^2_{\delta_2}}{n}) \end{bmatrix}^{-1} \begin{bmatrix} \frac{p-1}{p}\ \frac{b-1}{b}(1-\rho_a)\ \sigma^2_{a_1} \\ \frac{p-1}{p}\ \frac{b-1}{b}(1-\rho_a)\sigma_{a_{12}} \end{bmatrix}$$

$$\begin{bmatrix} b_5 \\ b_6 \end{bmatrix} = \begin{bmatrix} \frac{p-1}{p}\ (\sigma^2_{p_1} + \frac{\sigma^2_{e_1}}{b} + \frac{\sigma^2_{d_1}}{nb}) & \frac{p-1}{p}\ (\sigma_{p_{12}} + \frac{\sigma_{e_{12}}}{b} + \frac{\sigma_{\delta_{12}}}{nb}) \\ Sim. & \frac{p-1}{p}\ (\sigma^2_{p_2} + \frac{\sigma^2_{e_2}}{b} + \frac{\sigma^2_{\delta_2}}{nb}) \end{bmatrix}^{-1} \begin{bmatrix} \frac{p-1}{p}\ \frac{1+(nb-1)}{nb}\ \rho_a\ \sigma^2_{a_1} \\ \frac{p-1}{p}\ \frac{1+(nb-1)}{nb}\ \rho_a\ \sigma_{a_{12}} \end{bmatrix}$$

em que:

$\sigma^2_{\delta_i}$ *e* $\sigma_{\delta_{12}}$: variância dentro de parcelas para o caráter *i* e covariância dentro de parcela entre os caracteres 1 e 2, respectivamente.

$\sigma^2_{p_i}$ *e* $\sigma_{p_{12}}$: variância genética entre progênies para o caráter *i* e covariância genética ao nível de progênies entre os caracteres 1 e 2, respectivamente.

ρ_a : correlação genética entre os indivíduos do tipo de família considerada ($\rho_a = 0{,}50$, para famílias de irmãos germanos).

n, p e b : números de plantas por parcela, de progênies e de blocos, respectivamente;

Y e X : referem-se aos caracteres 1 e 2, respectivamente.

Para o caráter 2, um índice similar a I_1 pode ser construído, somente colocando o primeiro termo dos "multiplicandos" de b_1 a b_6 como função da covariância $\sigma_{a_{12}}$ e o segundo termo dos mesmos como função da variância $\sigma^2_{a_2}$.

Como exemplo, considere os caracteres produção de frutos (y) e altura da planta em cajueiros. A genealogia e os dados encontram-se a seguir.

Indivíduos	Bloco	Pai	Mãe	Produção de Frutos (kg)	Altura
4	1	1	-	22,7	3,0
5	2	3	2	14,6	3,2
6	2	1	2	19,7	2,5
7	1	4	5	17,7	2,8
8	1	3	6	25,3	3,5

Considere ainda os seguintes componentes de variância e parâmetros genéticos:

h^2_y = 0,33; h^2_x = 0,50; r_{axy} = 0,20; r_{xy} = 0,40;

$$G_O = \begin{matrix} \\ x \\ y \end{matrix}\begin{bmatrix} 0{,}10 & 0{,}103 \\ 0{,}103 & 2{,}75 \end{bmatrix} \qquad R_O = \begin{matrix} \\ x \\ y \end{matrix}\begin{bmatrix} 0{,}10 & 0{,}28 \\ 0{,}28 & 5{,}50 \end{bmatrix}$$

$$G_O^{-1} = \begin{matrix} \\ x \\ y \end{matrix}\begin{bmatrix} 10{,}4013 & -0{,}3896 \\ -0{,}3896 & 0{,}3782 \end{bmatrix} \qquad R_O^{-1} = \begin{matrix} \\ x \\ y \end{matrix}\begin{bmatrix} 11{,}6624 & -0{,}5937 \\ -0{,}5937 & 0{,}2120 \end{bmatrix}$$

As análises univariadas podem ser obtidas pelo mesmo sistema de equações, porém, fazendo-se:

$$G_O = \begin{bmatrix} 0{,}10 & 0 \\ 0 & 2{,}75 \end{bmatrix} \qquad R_O = \begin{bmatrix} 0{,}10 & 0 \\ 0 & 5{,}50 \end{bmatrix}$$

Os resultados, tanto para a análise multivariada quanto para as análises univariadas, são apresentados na seqüência.

Efeitos	Análise Multivariada		Análise Univariada	
	Altura da planta	Produção de frutos	Altura da planta	Produção de frutos
$\hat{b}_1$	3,1343	22,0184	3,1373	22,0274
$\hat{b}_2$	2,8344	17,1696	2,8365	17,1677
$\hat{a}_1$	-0,1353	0,5405	-0,1117	0,4980
$\hat{a}_2$	0,0036	-0,0901	-0,0004	-0,0908
$\hat{a}_3$	0,1831	-0,2587	0,1675	-0,2056
$\hat{a}_4$	-0,1448	-0,0173	-0,1390	-0,0534
$\hat{a}_5$	0,1808	-0,9966	0,1394	-0,9386
$\hat{a}_6$	-0,1496	0,9574	-0,1123	0,9033
$\hat{a}_7$	-0,0703	-1,2506	-0,1123	-1,2623
$\hat{a}_8$	0,1123	0,9129	0,1393	0,9336

As acurácias ($r_{\hat{a}a}$), tanto para a análise multivariada quanto univariada, foram obtidas a partir da expressão $r_{\hat{a}a}^2 = (\sigma_a^2 - PEV\ ) / \sigma_a^2$, em que PEV (para as equações de modelo misto em questão) é o elemento da diagonal da inversa da matriz dos coeficientes, pertencente a determinado indivíduo e caráter. Os valores de PEV e de acurácia são apresentados a seguir:

	Análise multivariada				**Análise univariada**			
Indivíduos	**PEV**		**Acurácia**		**PEV**		**Acurácia**	
	ALT.	**PROD.**	**ALT.**	**PROD.**	**ALT.**	**PROD.**	**ALT.**	**PC**
1	0,0909	2,5904	0,3011	0,2409	0,0911	2,5910	0,2977	0,2404
2	0,0973	2,7063	0,1631	0,1261	0,0974	2,7065	0,1606	0,1257
3	0,0863	2,5095	0,3702	0,2957	0,0866	2,5105	0,3659	0,2951
4	0,0773	2,3506	0,4766	0,3811	0,0778	2,3522	0,4711	0,3803
5	0,0775	2,3530	0,4739	0,3799	0,0780	2,3546	0,4686	0,3792
6	0,0814	2,4311	0,4314	0,3405	0,0819	2,4326	0,4257	0,3398
7	0,0811	2,4287	0,4344	0,3418	0,0816	2,4302	0,4285	0,3410
8	0,0759	2,3214	0,4906	0,3948	0,0764	2,3230	0,4853	0,3940

Verificam-se acurácias praticamente idênticas pelos métodos univariado e multivariado, fato que revela que, nas condições (parâmetros genéticos) estudadas, a seleção que emprega caráter auxiliar não traz vantagens. A eficiência da análise multivariada para o caráter altura, nesse caso, foi de apenas 1,011 (ou seja 1,1 %), valor esse que pode ser obtido pela equação da eficiência relatada anteriormente ou pela razão entre acurácia multivariada e univariada para o caráter altura na Tabela acima.

2.2 Índice de Seleção Multivariado

Os valores genéticos preditos ($\hat{a}$) pelo método multivariado podem ser utilizados na predição de valores genéticos de outros caracteres correlacionados a eles, por meio da expressão (Schneeberger et al., 1992):

$\hat{g} = G_{11}^{-1} G_{21} \hat{a}$, em que G_{11} é a matriz de variância-covariância genética entre os valores genéticos em $\hat{a}$; G_{21} é a matriz de covariância genética entre os valores genéticos de a e os de g em que g é o vetor de valores genéticos dos caracteres objetivos.

Também um índice de seleção pode ser calculado, incluindo informações econômicas dos vários caracteres. Nesse caso, tem-se $\hat{g} = (G_{11}^{-1} G_{21} w)\, \hat{a}$, em que $k = (G_{11}^{-1} G_{21} w)$ refere-se ao vetor de coeficientes de ponderação (k) dados aos vários caracteres e w refere-se aos pesos econômicos associados a tais caracteres. Quando os caracteres no índice e no objetivo ou agregado são os mesmos, tem-se $G_{21} = G_{11}$ e portanto $k = (G_{11}^{-1} G_{11} w) = Iw = w$ e $\hat{g} = w\ \hat{a}$. Assim, os ponderadores dos caracteres no índice são os próprios pesos econômicos e, para obtenção dos valores do índice de seleção, os valores genéticos preditos são ponderados diretamente pelos pesos econômicos. As herdabilidades, correlações genéticas e residuais e diferentes acurácias associadas aos vários caracteres já foram consideradas no processo de predição multivariada dos valores genéticos usados no índice.

A derivação do índice de seleção baseado em valores genéticos preditos é dada a seguir. Os ponderadores do tradicional índice de seleção baseado no fenótipo (vetor y) é dado por Var(y) k = Cov(a,y) w. Usando os valores genéticos preditos ($\hat{a}$) em lugar dos valores fenotípicos, tem-se que os ponderadores do índice são dados por $Var(\hat{a})$ k = $Cov(a, \hat{a})$ w. Como $Cov(a, \hat{a}) = Var(\hat{a})$, tem-se $Var(\hat{a})$ k = $Var(\hat{a})$ w, e portanto k = w.

3 TRANSFORMAÇÃO CANÔNICA E ÍNDICE DE SELEÇÃO VIA ANÁLISE UNIVARIADA

Ainda no contexto dos modelos multivariados, as variáveis correlacionadas podem ser transformadas em variáveis não correlacionadas, visando à realização de análises univariadas em vez de uma multivariada, porém sem perda de acurácia. Dentre os procedimentos de transformação, destacam-se: (i) transformação canônica, que é adequada quando todos os caracteres são avaliados em todos os indivíduos, ou seja, quando as matrizes de incidência X e Z são as mesmas para todos os caracteres e (ii) transformação Cholesky, quando observações de alguns caracteres são perdidas e a perda é seqüencial.

3.1 Transformação Canônica e Cholesky na Metodologia de Modelos Mistos

Devido à dificuldade de convergência e alto custo da análise sob modelo multivariado, o procedimento de transformação de variáveis originais em variáveis não correlacionadas geneticamente e residualmente, é muito interessante. Tal procedimento é denominado transformação canônica (Thompson, 1977b) e atua por meio da decomposição das matrizes de covariância genética e residual entre caracteres, usando matrizes de transformação.

Sendo $Var(y) = Var(a) + Var(e) = G + R$ e sendo G e R matrizes simétricas e positivas definidas, existe uma matriz de transformação T tal que TRT´= I e TGT´= H, em que I é uma matriz identidade e H é uma matriz diagonal. Assim, fazendo-se essas multiplicações, R se torna igual a identidade e G se torna diagonal, ou seja, não existem mais covariâncias explícitas em R e G. E multiplicando os dados originais por T, obtém-se o vetor y* das variáveis não correlacionadas.

Análises univariadas são então aplicadas sobre y* e os resultados (u) são convertidos para a escala original, produzindo valores genéticos idênticos àqueles obtidos sob o modelo multivariado (m). As expressões de conversão são dadas por $\hat{b}_m = T^{-1}\hat{b}_u^*$ e $\hat{a}_m = T^{-1}\hat{a}_u^*$. É importante mencionar que, para a realização das análises univariadas sobre y*, a herdabilidade é dada por $h_a^2 = \sigma_a^2/(\sigma_a^2 + 1)$ e o coeficiente lambda associado ao fator de *shrinkage* nas equações de modelo misto é dado por $\lambda_1 = 1/(\sigma_a^2)$, visto que a variável transformada apresenta variância residual unitária.

A matriz de transformação T é obtida da maneira relatada a seguir. Tomando-se as matrizes G e R apresentadas anteriormente, tem-se os seguintes passos.

a) Obtenção das matrizes G e R:

$$G = \begin{matrix} x \\ y \end{matrix}\begin{bmatrix} 0{,}10 & 0{,}103 \\ 0{,}103 & 2{,}75 \end{bmatrix} \qquad R = \begin{matrix} x \\ y \end{matrix}\begin{bmatrix} 0{,}10 & 0{,}28 \\ 0{,}28 & 5{,}50 \end{bmatrix}$$

Uma aproximação para obter G e R pode basear-se em análises univariadas sobre as variáveis originais e posterior cômputo das correlações entre valores genéticos e entre resíduos das diferentes variáveis para obtenção de G e R, respectivamente.

b) Decomposição de R em termos das matrizes de autovalores (V) e autovetores (W):

R = W V W´

Usando G e R acima tem-se:

$$V = \begin{bmatrix} 0{,}0855 & 0 \\ 0 & 5{,}5145 \end{bmatrix} \qquad W = \begin{bmatrix} 0.9987 & 0{,}0516 \\ -0{,}0516 & 0{,}9987 \end{bmatrix}$$

c) Cálculo da matriz P e de PGP´:

P = W $(V^{-1})^{1/2}$ W´

Usando G, W e V acima, tem-se:

$$P = \begin{bmatrix} 3{,}4115 & -0.1544 \\ -0.1544 & 0{,}4338 \end{bmatrix} \qquad PGP' = \begin{bmatrix} 1{,}1209 & -0{,}0820 \\ -0{,}0820 & 0{,}5061 \end{bmatrix}$$

d) Decomposição de PGP´ em termos das matrizes de autovalores (Z) e autovetores (Q):

PGP´ = QZQ´

Usando PGP´ acima, tem-se:

$$Z = \begin{bmatrix} 1{,}1317 & 0 \\ 0 & 0{,}4954 \end{bmatrix} \qquad Q = \begin{bmatrix} 0.9915 & 0{,}1300 \\ -0{,}1300 & 0{,}9915 \end{bmatrix}$$

e) Obtenção da matriz de transformação T:

Usando Q e P acima, tem-se:

$$T = Q'P = \begin{bmatrix} 3{,}4027 & -0{,}2095 \\ 0{,}2902 & 0{,}4101 \end{bmatrix} \qquad T^{-1} = \begin{bmatrix} 0{,}2816 & 0{,}1438 \\ -0{,}1993 & 2{,}3367 \end{bmatrix}$$

A transformação dos efeitos genéticos para a escala original é ilustrada a seguir usando os resultados para o indivíduo 1, apresentados anteriormente.

Efeitos	Análise Multivariada		Análise Univariada após transformação canônica	
	Altura da planta	Produção de frutos	Altura da planta	Produção de frutos
$\hat{a}_1$	-0,1353	0,5405	-0,5736	0,1824

$$\hat{a}_m = T^{-1}\hat{a}_u^*$$

$$\begin{bmatrix} \hat{a}_{11m} \\ \hat{a}_{12m} \end{bmatrix} = \begin{bmatrix} 0{,}2816 & 0{,}1438 \\ -0{,}1993 & 2{,}3367 \end{bmatrix} \begin{bmatrix} \hat{a}_{11u}^* \\ \hat{a}_{12u}^* \end{bmatrix} = \begin{bmatrix} 0{,}2816 & 0{,}1438 \\ -0{,}1993 & 2{,}3367 \end{bmatrix} \begin{bmatrix} -0{,}5736 \\ 0{,}1824 \end{bmatrix} = \begin{bmatrix} -0{,}1353 \\ 0.5405 \end{bmatrix}$$

Verifica-se que os mesmos resultados da análise multivariada foram obtidos.

Outro tipo de transformação que pode ser usada é a transformação Cholesky. Tal transformação produz variáveis com correlação residual nula e variância unitária. A matriz de transformação L^{-1} é obtida por meio da decomposição Cholesky de R. Tal decomposição é dada por R = LL´, em que L é uma matriz triangular inferior. A matriz de transformação é a inversa de L.

A decomposição Cholesky é uma especialização da decomposição LU (em matriz triangular inferior L e superior U), a qual é uma conseqüência do método de eliminação de Gauss ou de absorção de uma linha da matriz por vez, via operações elementares sobre linhas da matriz, em processo que culmina com a triangularização. No processo de eliminação de Gauss, os elementos diagonais são denominados pivôs. A transformação de variáveis via decomposição Cholesky é equivalente ao processo de condensação pivotal (Aitken, 1937) usado na obtenção de variáveis canônicas.

3.2 Índice de Seleção via Transformação Canônica

O índice de seleção tradicional de Hazel (1943) é baseado nos valores fenotípicos y e é dado por $I = \sum_{i=1}^{n} k_i \; y_i$, em que os ponderadores k são estimados, levando-se em conta as informações econômicas, as herdabilidades e as correlações genéticas e fenotípicas. Tais ponderadores são estimados por $k = (F_{11}^{-1} G_{11} w)$, em que F_{11} é a matriz de covariâncias fenotípicas entre os caracteres.

Esse índice torna-se genético pois, apesar de ponderar valores fenotípicos, usa as herdabilidades dos vários caracteres. Capitaliza também os benefícios da seleção com caracteres auxiliares, já que considera também as correlações genéticas e fenotípicas entre caracteres. Assim, se existem grandes diferenças entre as correlações genéticas e ambientais ou fenotípicas, haverá maior acurácia seletiva dos caracteres, conforme mostrado com detalhes anteriormente. No entanto, tal índice não corrige os valores fenotípicos para os efeitos ambientais identificáveis e também não considera a correlação genética entre indivíduos aparentados. Por isso, é um índice inferior ao índice baseado em valores genéticos preditos pelo procedimento BLUP multivariado.

O índice de Hazel pode adotar as seguintes simplificações:

- Com pesos econômicos e herdabilidades iguais e correlações nulas: $I = \sum_{i=1}^{n} y_i \; \frac{1}{\sigma_{y_i}}$, em que σ_{y_i} é o desvio padrão fenotípico da característica y_i.
- Com pesos econômicos iguais e correlações nulas: $I = \sum_{i=1}^{n} h_i^2 \; y_i \; \frac{1}{\sigma_{y_i}}$.
- Com herdabilidades iguais e correlações nulas: $I = \sum_{i=1}^{n} w_i \; y_i$ (Índice Base, de Williams, 1962).
- Com correlações nulas: $I = \sum_{i=1}^{n} w_i \; h_i^2 \; y_i$ (Índice Primário, de Andrus e Mc Gilliard, 1975).
- Com valores genéticos preditos de forma independente para cada caráter e, posteriormente, ponderados por seus valores econômicos relativos, tem-se o Índice de Seleção Aproximado (Van Vleck et al., 1987). O índice, neste caso, fornece resultados similares aos do Índice Primário.

O índice de Hazel é, em teoria, o mais acurado. Na prática, ele será realmente melhor se as estimativas dos parâmetros genéticos forem confiáveis. Em situações em que isto não acontece, os índices simplificados podem conduzir a melhores resultados. Um caso típico é o do Índice de Seleção Aproximado, o qual pode ser usado com vantagens se: (i) as estimativas de correlações genéticas não são confiáveis; (ii) as correlações genéticas e fenotípicas são próximas de zero; (iii) as correlações genética e ambiental apresentam magnitudes similares; (iv) as acurácias na predição dos valores genéticos para cada caráter individualmente são altas (próximas de 1, tal como ocorre com a seleção de genitores com base em teste de suas progênies). Neste caso, o uso de outros caracteres correlacionados praticamente não contribui para aumentar a acurácia seletiva.

Esse índice de seleção aproximado é calculado a partir de valores genéticos preditos por BLUP univariado, o qual considera as herdabilidades dos caracteres, corrige os valores fenotípicos para os efeitos ambientais identificáveis e também considera a correlação genética entre indivíduos aparentados. Por isso, pode ser superior ao índice de Hazel. No entanto, não considera as diferenças entre as correlações genéticas e fenotípicas e, portanto, pode ser um índice inferior ao índice baseado em valores genéticos preditos pelo procedimento BLUP multivariado, se essas diferenças entre correlações forem de grandes magnitudes.

A única alternativa de construir um índice de seleção tão eficiente quanto aquele obtido sob BLUP multivariado, porém usando valores genéticos preditos por análise univariada, é via transformação canônica das variáveis originais. Essa transformação produz variáveis transformadas y* com correlações genéticas e residuais nulas. Após análises univariadas de y*, a matriz de transformação T é aplicada sobre os valores genéticos preditos ($\hat{a}_u$), produzindo novos valores, idênticos aos obtidos sob análise mutivariada. Esses valores genéticos ($\hat{a}_m$) são então ponderados por seus pesos econômicos, gerando o índice de seleção de eficiência máxima.

Após análises univariadas de y*, a seqüência de cálculos é então:

(i) $\hat{a}_m = T^{-1}\hat{a}_u$

(ii) $\hat{g} = w\ \hat{a}_m$

Uma alternativa eficiente para cômputo dos pesos econômicos w_i refere-se ao uso das correlações genéticas entre cada caráter i e o caráter objetivo do melhoramento j (r_{gij}). Nesse caso, w_i é dado por $w_i = r_{gij} / \sum_{i=1}^{n} r_{gij}$, ou seja, equivale à correlação como proporção do somatório das correlações envolvendo as n variáveis e o caráter objetivo.

4 DISTÂNCIAS ESTATÍSTICAS E ANÁLISE DE AGRUPAMENTO

Em algumas situações no melhoramento, a inferência sobre a divergência genética dos genitores a serem usados em cruzamento pode ser relevante. Tal inferência pode se basear em divergência filogenética (entre espécies diferentes), divergência geográfica, informação de genealogia (coeficiente de parentesco), capacidade específica de combinação (CEC) para o caráter de interesse obtidas via cruzamentos dialélicos, agrupamento baseado em distâncias multivariadas, dispersão gráfica após análise multivariada de caracteres múltiplos.

Um procedimento ótimo para análise estatística da divergência genética baseada em caracteres múltiplos deve considerar as seguintes premissas: (i) assumir os efeitos genotípicos como aleatórios; (ii) basear-se nos valores genotípicos e não nos fenotípicos; (iii) considerar o desbalanceamento dos dados; (iv) realizar a análise da divergência simultaneamente à predição dos valores genéticos, pois os valores genotípicos são preditos com diferentes precisões e isto precisa ser considerado no método de análise. Este último aspecto demanda a união das técnicas de análise multivariada (envolvendo vários caracteres) e de modelos mistos (REML/BLUP). A análise de agrupamento com base nas CEC para o caráter de interesse, obtidas sob REML/BLUP, também é um procedimento adequado.

Dentre as medidas estatísticas de distância mais utilizadas, citam-se a distância euclidiana e a distância de Mahalanobis. Uma breve descrição dessas medidas é apresentada a seguir.

4.1 Distância Euclidiana ou Métrica Euclidiana

A distância euclidiana ou métrica euclidiana equivale à raiz quadrada do somatório da diferença quadrática entre os valores de cada variável observada em dois indivíduos ou pontos de um espaço v – dimensional, em que v é o número de variáveis ou coordenadas no espaço. Matematicamente é dada por $D_E(i,i') = [\sum_{j=1}^{v}(u_{ij} - u_{i'j})^2]^{1/2}$, em que i e i´ referem-se aos dois indivíduos em questão e u_{ij} e $u_{i'j}$ referem-se aos valores observados para a variável j nos indivíduos i e i´.

Os valores obtidos por essa expressão são muito altos quando o número de variáveis é elevado. Em função disso, formas alternativas dessa distância são mais usadas na prática tais quais a distância euclidiana média dada por $D_{EM}(i,i') = [\sum_{j=1}^{v}(u_{ij} - u_{i'j})^2 / v]^{1/2}$ e o quadrado da distância euclidiana média dado por $D_{QEM}(i,i') = (\sum_{j=1}^{v}(u_{ij} - u_{i'j})^2) / v$.

A distância euclidiana não é invariante aos efeitos de escala de medição das variáveis. Assim, as variáveis usadas em u_{ij} e $u_{i'j}$ devem ser previamente padronizadas, dividindo-as por seus respectivos desvios padrões.

4.2 Distância Estatística ou de Mahalanobis

A distância generalizada de Mahalanobis ou distância estatística foi proposta em 1936 (Mahalanobis, 1936) e difere da distância euclidiana por considerar as correlações entre as variáveis e ser invariante aos efeitos de escala de medição dos caracteres. Dados os vetores de médias de v variáveis para o indivíduo i, $u_i = (u_{i1}, u_{i2}, u_{i3....}, u_{iv})'$ e para o indivíduo i´, $u_{i'} = (u_{i'1}, u_{i'2}, u_{i'3....}, u_{i'v})$ bem como a matriz de covariâncias Σ entre as variáveis, a distância de Mahalanobis é dada por $D_M(i,i') = [(u_i - u_{i'})'\Sigma^{-1}(u_i - u_{i'})]^{1/2}$.

Se a matriz de covariâncias é uma matriz identidade (caso de variáveis não correlacionadas e com variâncias unitárias), a distância de Mahalanobis reduz-se à distância Euclidiana. Se a matriz de covariâncias é uma matriz diagonal (caso de variáveis não correlacionadas), a distância de Mahalanobis reduz-se à distância Euclidiana normalizada ou padronizada dada por $D_E(i,i') = [\sum_{j=1}^{v} (u_{ij} - u_{i'j})^2 / \sigma_j^2]^{1/2}$. A distância de Mahalanobis é também intimamente relacionada à estatística T^2 de Hotelling usada para comparação multivariada de médias. O uso da distância de Mahalanobis é preferível em relação ao uso da distância euclidiana, pois a primeira considera as correlações entre as variáveis.

4.3 Agrupamento pelo Método de Tocher

Técnicas de agrupamento visam separar indivíduos ou entidades em grupos heterogêneos entre eles e homogêneos dentro deles. Várias técnicas de agrupamento existem, as quais podem ser subdivididas em duas classes: métodos hierárquicos e métodos mutuamente exclusivos. Os métodos hierárquicos conduzem a dendogramas e os métodos mutuamente exclusivos conduzem a grupos distintos não conectados entre si. Assim, esses últimos são de interpretação mais fácil e imediata. Dentre esses, destaca-se o método de agrupamento de Tocher (Rao, 1952). Tal método tem grande uso no melhoramento de plantas e é descrito em detalhes por Cruz e Regazzi (1994). Esse método adota o critério de que a média das medidas de distância dentro de cada grupo deve ser menor que as distâncias médias entre grupos. Inicialmente, o par de indivíduos mais próximos é identificado e então é formado o primeiro grupo. Em seguida, aplica-se o critério mencionado acima, visando verificar se novos indivíduos possam ser alocados no mesmo grupo. Caso não possam ser incluídos em um grupo, novos grupos são formados.

A distância média intragrupo aumenta com a entrada de um novo indivíduo no grupo. Para aceitar a entrada desse novo indivíduo no grupo, esse acréscimo na distância média intragrupo ($D_{(ij)k}/n$) deve ser comparado com um limite máximo permitido para inclusão no grupo.

Esse limite máximo (max.) é geralmente tomado como a maior distância dentre todas as distâncias mínimas associadas a cada indivíduo.

Matematicamente, isto é expresso por:

a) inclusão do indivíduo k no grupo: se $D_{(ij)k}/n \leq$ max.

b) não inclusão do indivíduo k no grupo: se $D_{(ij)k}/n >$ max.

A quantidade n refere-se ao número de indivíduos já alocados no grupo e $D_{(ij)k}$ é dado por $D_{(ij)k} = D_{ik} + D_{jk}$.

O método de Tocher pode ser aplicado sobre várias medidas de distância, tais como funções das capacidades específicas de combinação, complementos de correlações genéticas entre locais, dentre outras.

5 ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais (PCA) é uma técnica de transformação linear para redução de dimensionalidade em conjunto de dados multivariados. É também denominada transformação de Hotelling. Apresenta vários usos na área experimental tais quais descarte de variáveis redundantes, dispersão gráfica e agrupamento de indivíduos ou genótipos, geração de novas variáveis ou componentes para uso em índice de seleção, seleção multivariada no contexto dos modelos mistos.

5.1 Componentes Principais Tradicional

Os componentes principais de um conjunto de v variáveis correlacionadas são novas v variáveis com as seguintes propriedades: (i) são funções lineares das v variáveis originais; (ii) são não correlacionadas umas com as outras; (iii) explicam sucessivamente o máximo da variação original.

A matriz de covariância das v variáveis é dada por $\sum = \Lambda\ \Lambda' = VD_\alpha V'$, em que D_α é a matriz diagonal dos autovalores ou raízes características, V é a matriz dos autovetores e $\Lambda = V(D_\alpha)^{1/2}$. A partir dessa decomposição da matriz de covariância, os componentes principais são dados por $PC_i = v_i y$, em que y é vetor das variáveis originais, v_i é o autovetor i associado ao autovalor i, com $v_i' v_i = 1$. Assim, os pesos das variáveis em cada componente principal é dado pelos elementos do autovetor correspondente e o somatório do quadrado desses pesos para um componente principal equivale a 1. A variância de cada PC é dada pelo correspondente autovalor.

Ordenando os autovalores em ordem decrescente, tem-se que o primeiro PC associado ao primeiro autovalor ordenado e respectivo autovetor explica a maior parte da variação total, seguido pelo segundo e assim sucessivamente. Para um determinado número de termos ou componentes, m $<$ v, grande parte (80 % ou mais) da variação total é explicada por um pequeno número das novas variáveis. PC's com autovalores próximos a zero não contribuem com informação adicional àquela já retida nos primeiros componentes principais e podem ser descartados sem perda de informação, e a dimensionalidade dos dados é reduzida. Considerando apenas os primeiros m componentes principais, a matriz de covariância das variáveis é dada por $\sum^* = \Lambda_m\ \Lambda_m' = V_m D_{\alpha m} V_m'$, em que m indica a dimensão dessas matrizes em termos de colunas. A matriz $\sum^*$, de tamanho v x v, tem posto igual a m.

Na PCA, a seguinte seqüência de cálculos é empregada, usando dados de um exemplo hipotético.

a) Obtenção das matrizes de médias e de correlações.

Considere o seguinte conjunto de dados associados à avaliação de seis variáveis em dez indivíduos.

Indivíduo	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
1	3994	2282	3068	1694	4320	1382
2	5316	1981	4168	3470	5900	2214
3	5050	1883	4408	3373	6476	1761
4	5996	2162	3614	3003	5021	1411
5	6086	5601	4599	3091	6145	2014
6	5180	2506	5022	2506	8299	2478
7	5291	1998	5241	2614	7768	2247
8	6148	2444	6147	2278	7440	2026
9	7292	2784	5692	2592	7978	2586
10	6615	3204	7655	3331	9225	3336

A matriz de correlação calculada equivale a:

Variáveis	1	2	3	4	5	6
1	1	0.3645	0.6681	0.3368	0.5599	0.5738
2	0.3645	1	0.1827	0.1468	0.0784	0.1961
3	0.6681	0.1827	1	0.2176	0.9099	0.8786
4	0.3368	0.1468	0.2176	1	0.191	0.3295
5	0.5599	0.0784	0.9099	0.191	1	0.9003
6	0.5738	0.1961	0.8786	0.3295	0.9003	1

b) Cálculo dos autovalores e autovetores.

As estimativas dos autovalores são obtidas por $\det(R-\alpha_i I)=0$ e dos autovetores são obtidas pela solução do sistema $(R-\alpha_i I)v_i=\Phi$, em que R é a matriz de correlações entre as variáveis, α_i refere-se ao autovalor i, v_i é o autovetor i e Φ é um vetor nulo.

Os seguintes autovalores foram obtidos:

Autovalores	3.458	1.0933	0.8507	0.429	0.0973	0.0716

Os autovetores calculados são apresentados na Tabela seguinte:

Autovetor 1	Autovetor 2	Autovetor 3	Autovetor 4	Autovetor 5	Autovetor 6
0.4209	-0.2618	-0.0762	-0.8412	0.1937	0.0583
0.1687	-0.7488	-0.5238	0.3559	-0.0625	0.0767
0.5064	0.1977	-0.1029	0.0123	-0.6722	-0.4917
0.2178	-0.4528	0.8403	0.1469	-0.1317	0.0502
0.4892	0.3191	-0.0489	0.1813	-0.0998	0.7833
0.5011	0.1574	0.0269	0.3335	0.6923	-0.3644

Os autovetores apresentam três usos principais: (i) reconstrução da matriz de correlação com base no modelo reduzido; (ii) obtenção dos escores dos componentes principais para todos os indivíduos; (iii) descarte de variáveis redundantes: os maiores coeficientes das variáveis nos últimos componentes principais (aqueles com autovalores menores que 0.70) podem ser usados para o descarte de variáveis redundantes que pouco contribuem para a discriminação dos genótipos.

c) Cálculo da proporção acumulada da variação total explicada pelos componentes principais e respectivos autovalores.

A importância relativa de cada componente principal é dada pela proporção da variação total que ele explica, isto é, $p = Var(PC_i) / \sum_i Var(PC_i) = \alpha_i / \sum_i \alpha_i = \alpha_i / tr(R) = \alpha_i / v$. A proporção acumulada explicada pelos PC's é dada pela soma dos respectivos p's. Os resultados do presente exemplo são apresentados a seguir.

Ordem	Autovalores	Proporção Explicada	Proporção Explicada Acumulada
1	3.46	0.5763	0.5763
2	1.09	0.1822	0.7586
3	0.85	0.1418	0.9003
4	0.43	0.0715	0.9718
5	0.10	0.0162	0.9881
6	0.07	0.0119	1

Verifica-se que os dois primeiros componentes principais explicam cerca de 76 % da variação total. Assim, inferências baseadas nesses dois são adequadas na prática, seja para dispersão gráfica dos escores associados aos indivíduos, em dois eixos cartesianos (componente 1 na abcissa e componente 2 na ordenada), seja para cálculo da distância euclidiana entre indivíduos usando as duas novas variáveis (componentes), ou para nova análise univariada das mesmas. Os três últimos autovalores são menores que 0.70 e, então, os maiores elementos de cada um dos autovetores associados aos três últimos componentes principais podem ser usados para a identificação e descarte de variáveis redundantes.

d) Determinação dos escores dos componentes principais associados a todos os indivíduos da análise.

Os escores ou novas variáveis (vetor z) associados aos indivíduos são dados pelos produtos dos autovetores pelo vetor das variáveis originais ou padronizadas (y_i) associadas a cada indivíduo, ou seja, $z_i = v_i^{'} y_i$.

Para o presente exemplo, os escores (associados às variáveis padronizadas) dos componentes principais são apresentados na seqüência:

Indivíduos	**Escores no PC1**	**Escores no PC2**	**Escores no PC3**	**Escores no PC4**	**Escores no PC5**	**Escores no PC6**
1	-3.45	0.63	-1.12	0.37	0.14	-0.34
2	-0.57	-0.26	1.47	0.21	0.35	-0.25
3	-0.86	-0.00	1.34	0.21	-0.38	0.21
4	-1.59	-0.68	0.66	-1.04	-0.03	0.03
5	0.26	-2.57	-0.93	0.50	-0.08	0.11
6	0.39	0.90	-0.34	0.70	0.24	0.43
7	0.12	0.98	0.04	0.27	-0.08	0.20
8	0.50	0.72	-0.81	-0.61	-0.55	-0.10
9	1.68	0.11	-0.55	-1.08	0.46	0.11
10	3.51	0.17	0.25	0.46	-0.07	-0.41

Os escores dos primeiros componentes principais associados aos indivíduos podem ser usados para agrupamento de indivíduos via dispersão gráfica e também via o método de Tocher aplicado sobre as distâncias euclidianas médias calculadas sobre esses escores. Esses escores podem também ser submetidos a análises univariadas para posterior uso dos resultados em índices de seleção.

e) Cálculo das correlações entre as variáveis e os componentes principais.

Quando se usa a matriz de correlação entre as variáveis, as correlações entre as variáveis e os componentes principais são estimadas por $r_{vi} = v'_{vi}(\alpha_i)^{1/2}$, em que v_{vi} é o ponderador da variável v no componente principal i.

Para o presente exemplo, as correlações entre as variáveis e os componentes principais são apresentados na seqüência:

Componente	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
1	0.78	0.31	0.94	0.41	0.91	0.93
2	-0.27	-0.78	0.21	-0.47	0.33	0.16
3	-0.07	-0.48	-0.09	0.78	-0.05	0.02
4	-0.55	0.23	0.01	0.10	0.12	0.22
5	0.06	-0.02	-0.21	-0.04	-0.03	0.22
6	0.02	0.02	-0.13	0.01	0.21	-0.10

f) Dispersão gráfica dos escores associados aos 2 ou 3 primeiros componentes principais.

Um exemplo completo e real de PCA é discutido com detalhes no Capítulo 10.

5.2 Componentes Principais sob Modelos Mistos (PCAM)

O método PCA descrito no item anterior geralmente baseia-se em dados fenotípicos e correlações fenotípicas. As inferências baseadas nesse tipo de análise podem ser pobres do ponto de vista genético. Uma alternativa é trabalhar com os valores genéticos preditos e com as correlações genéticas, conforme realizado pelo *software* Selegen-Reml/Blup. Outra alternativa é reunir as técnicas de análise multivariada e de modelos mistos e produzir uma análise direta e em um só passo, no nível genético. Esta análise simultânea tem grande aplicação na análise de múltiplos caracteres e de medidas repetidas.

A metodologia de modelos mistos padrão pode ser usada para estimar autovalores e autovetores diretamente sem a necessidade de se estimar a matriz de covariância (Σ) completa. A principal diferença para o modelo multivariado misto tradicional refere-se ao fato de que os parâmetros a serem estimados fazem parte da matriz de incidência dos efeitos genéticos aleatórios, conduzindo à estimação sob posto reduzido.

Outra vantagem dessa abordagem refere-se ao fato de que a estimação direta da estrutura de covariância garante que a matriz de covariância será positiva definida, fato que não é garantido por outros métodos de estimação de $\sum$. Assim, a inclusão de caracteres adicionais na análise contribui para aumentar a precisão na estimação ao invés de desestabilizar as estimativas. Também PCA's podem ser estimados tanto no nível genético quanto ambiental, desdobrando a tradicional PCA fenotípica.

A seguir, é apresentada uma extensão dos modelos mistos para incorporar a análise de componentes principais com base em Meyer e Kirkpatrick (2005).

Modelo Misto Tradicional

$y = Xb + Za + e$

PCA sob Modelo Misto (PCAM)

$y = Xb + Z(Q \otimes I_g)(Q^{-1} \otimes I_g)a + \varepsilon = Xb + Z^* a^* + \varepsilon$, em que:

$Q = V_m$ e $a_j^* = Q' a_j$.

Os valores genéticos do indivíduo j para os caracteres originais é dado por $\hat{a}_j = Q\hat{a}_j^*$. I_g é a matriz identidade com ordem igual ao número g de genótipos.

Sob esse modelo, a matriz de covariância genética é dada por $\sum = \Lambda\ \Lambda'$, em que $\Lambda\ \Lambda'$= VD_α V´ , D_α é a matriz diagonal dos m autovalores e V é a matriz dos autovetores. Escolhendo-se V e D_α referentes apenas à dimensão m, esse modelo misto é reduzido e ajusta somente os m primeiros componentes principais. Assim, na técnica PCAM, a estrutura de covariância é simplificada para $\sum^* = \Lambda_m\ \Lambda_m' = V_m D_{\alpha m} V_m'$ em que m indica uma das dimensões dessas matrizes (número de colunas).

O método REML é então aplicado para estimação nesse modelo. Escolhendo-se $Q = V_m (D_{\alpha m})^{1/2}$ permite-se que os elementos de V_m, os quais são determinados por restrições de ortogonalidade em V, possam ser obtidos por meio da solução de um pequeno sistema de equações lineares. Estimativas de α_i são então obtidas via cálculo das normas das respectivas colunas de Q. Isto conduz a $Var(a^*) = I_{mg}$ e o logaritmo da função de verossimilhança restrita a ser maximizada é dado por $-2LogL = const. + \log|A| + \log|R| + \log|C| + y'Py$, em que $R = Var(e)$, C é a matriz dos coeficientes das equações de modelo misto sob o modelo PCAM, A é a matriz de correlação genética aditiva entre os indivíduos em avaliação e $y'Py$ é a soma de quadrados dos resíduos ponderada. A verossimilhança é invariante a transformações ortogonais aplicadas a Q. Considerando a decomposição Cholesky $\sum = L\ L'$, a decomposição de valor singular da fatoração Cholesky produz $L = V(D_\alpha)^{1/2} T'$, com T´T = I. Nesse caso, os vetores singulares a esquerda de L são iguais aos autovetores de $\sum$ e os valores singulares são iguais a $(\alpha)^{1/2}$, conforme Harville (1997). Assim, escolhendo T´ como transformação, produz-se Q = L e as estimativas dos primeiros m autovetores e autovalores da matriz de covariância podem ser obtidas pela estimação dos elementos não zero das m primeiras colunas da fatoração Cholesky de $\sum$. Fazendo-se a rotação do espaço paramétrico dessa forma, automaticamente considera-se a restrição no número de parâmetros. Maiores detalhes sobre essa rotação são apresentados no Capítulo 8 em associação com a técnica FAMM.

A técnica PCAM surge como uma terceira classe de métodos adequados ao estudo de dados longitudinais ou medidas repetidas. A primeira dessas classes é não paramétrica e refere-se ao uso de funções de covariância baseadas em funções flexíveis tais quais polinômios, gerando o método de regressão aleatória. Essa classe pode basear-se também em outras funções tais quais as splines cúbicas e splines tipo B, as quais apresentam melhor comportamento do que os polinômios mas requerem o ajuste de um maior número de parâmetros. A segunda classe de métodos usa formas paramétricas simples para a função de covariância, gerando os métodos de processo caráter e de antedependência (ver Capítulo 9). Essa abordagem reduz o número de parâmetros e, portanto, torna os resultados mais precisos e confiáveis, pois menos erros de estimação são introduzidos. A terceira classe, na qual a técnica PCAM se insere, envolve a estimação direta e é vantajosa sobre as duas primeiras porque não assume *a priori* qualquer forma para a função de covariância e envolve menos parâmetros do que os métodos de regressão aleatória. Assim, é uma boa alternativa para os casos em que o uso de uma determinada forma paramétrica simples (por exemplo, estruturas ARH e SAD, descritas no Capítulo 9) não apresentar uma justificativa biológica.

6 ANÁLISE DE FATORES

Conforme relatado anteriormente, as técnicas PCA e FA são relacionadas. No entanto, diferenças conceituais muito importantes existem entre as duas técnicas. Na AF, os fatores são variáveis latentes não observáveis e na PCA os componentes não são variáveis latentes e são combinações lineares de variáveis observadas. Se o objetivo da análise for a redução de dimensionalidade dos dados, a PCA deve ser usada. Se o objetivo for a construção de um modelo testável que explique as inter-relações entre as variáveis, a AF deve ser usada. Outras diferenças e similaridades entre essas técnicas são descritas nos tópicos seguintes.

6.1 Análise de Fatores Tradicional

A análise de fatores (AF), assim como as técnicas de componentes principais (PCA) e componentes principais canônicos (ou componentes principais via variáveis canônicas - PCAC), é uma técnica de análise multivariada utilizada no estudo da estrutura de correlação ou covariância envolvendo várias variáveis simultaneamente. O objetivo principal é a simplificação ou explicação desta estrutura de correlação entre um grande número de variáveis por meio de poucas variáveis aleatórias não observáveis denominadas fatores. De forma similar, as técnicas PCA e PCAC visam resumir as informações em poucos componentes independentes entre si e que retenham o máximo possível da variação total original. Estas três técnicas consideram as correlações entre as variáveis e geram um novo conjunto de variáveis independentes. A técnica PCA baseia-se em apenas uma observação por grupo ou indivíduo em cada variável, ou seja, utiliza apenas uma matriz de médias de variáveis nos vários indivíduos. Por outro lado, a técnica PCAC faz uso da matriz de médias e da matriz de dispersão (covariância) residual das variáveis, sendo, portanto, aplicada quando se dispõe de informações de repetições ou de vários indivíduos dentro dos grupos ou objetos.

Estas técnicas relatadas são apresentadas em detalhes em várias obras (Johnson e Wichern, 1982; Mardia et al., 1988; Souza, 1988; Duarte e Vencosky, 1999; Cruz e Regazzi, 1994; Cruz e Carneiro, 2003). A descrição apresentada neste documento segue de perto o material apresentado por Johnson e Wichern (1982).

A AF tem como origem os trabalhos de Karl Pearson, em 1901, e Charles Spearman, em 1904. Ambos trabalharam na Inglaterra, Spearman foi Estatístico e Professor de Psicologia no *University College of London* (*UCL),* e o Matemático Pearson foi o fundador da Estatística e Biometria trabalhando também no UCL – Departamento de Estatística Aplicada. A AF pode ser considerada como uma extensão da PCA. Entretanto, para o modelo fatorial são feitas algumas suposições básicas que não são exigidas para componentes principais.

Várias técnicas estatísticas são relacionadas à análise de componentes principais, tais quais: PCAC, AF, modelos mistos fatores analíticos (FAMM), modelos aditivos para os efeitos principais e multiplicativos para a interação (AMMI, também denominado análise de componentes principais

centrada). Além do estudo e simplificação da estrutura de correlações, estas técnicas permitem a dispersão dos indivíduos em gráficos, empregando os primeiros eixos principais, denominados componentes ou fatores. Assim são úteis no agrupamento de variáveis e também na subdivisão dos indivíduos em grupos similares.

6.1.1 Modelo Fatorial Ortogonal

Considere que as variáveis observadas são agrupadas de acordo com suas correlações. Um fator é representado por um grupo de variáveis altamente correlacionadas entre si, mas, com correlações relativamente baixas com as variáveis de um outro grupo de variáveis.

Considere o vetor observável y, com p variáveis componentes, em que $y \sim (\mu, \sum)$. O modelo fatorial postula que y é linearmente dependente sobre algumas variáveis aleatórias não observáveis ou latentes $f_1, f_2, ..., f_m$ (com $m \leq p$) denominadas **fatores comuns** e p fontes de variação aditivas $\delta_1, \delta_2, ..., \delta_p$ denominados erros ou **fatores específicos**.

O modelo fatorial ortogonal é dado por:

$$y_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + ... + \lambda_{1m} f_m + \delta_1$$
$$y_2 - \mu_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + ... + \lambda_{2m} f_m + \delta_2$$
$$\vdots$$
$$y_p - \mu_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + ... + \lambda_{pm} f_m + \delta_3$$

em que:

μ_i: média da i-ésima variável y_i;

f_j: j-ésimo fator comum a todas as variáveis;

λ_{ij}: peso ou carregamento do j-ésimo fator f_j na i-ésima variável y_i;

i = 1, 2, ..., p;

j = 1, 2, ..., m.

Assim, cada variável y_i é decomposta em termos de m fatores comuns a todas as variáveis e um fator específico ou erro típico da própria variável. Pesos diferentes (λ_{ij}) são dados aos fatores comuns, no modelo que explica cada uma das variáveis.

Em notação matricial tem-se:

$Y - \mu = \Lambda f + \delta$, em que:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}; \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}; \qquad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix}$$

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}; \qquad \Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix}$$

O vetor $(y - \mu)$ é expresso em termos de p + m variáveis aleatórias (m: fatores comuns e p: fatores específicos) as quais não são observáveis. Este fato distingue o modelo fatorial do modelo de regressão multivariada (regressão fixa) no qual as variáveis independentes (análogas aos fatores f_j no modelo fatorial) podem ser observadas. A exigência $m \leq p$ para o número de fatores implica que a estrutura de correlação via uso do modelo fatorial não é mais complicada do que a estrutura original. E quando m é pequeno em relação à p, a análise de fatores é mais útil.

Tem-se as seguintes suposições adicionais em relação aos vetores aleatórios f e δ, em termos de estrutura de média e variância:

$E(f) = 0; \quad E(\delta) = 0$.

$Var(f) = E(f\ f') = I$; pois os fatores comuns possuem variância unitária e covariância nula.

$Var(\delta) = E(\delta\ \delta') = \psi = diag_{(\psi_1 \ldots \psi_p)}$, em que $\psi_i = Var(\delta)$.

$COV(f, \delta) = 0$, ou seja, f e δ são independentes.

Com estas suposições, o modelo fatorial descrito anteriormente constitui o modelo fatorial ortogonal. A partir de tal modelo, obtém-se:

$$\begin{aligned}(y-\mu)(y-\mu)' &= (\Lambda\ f+\delta)\ (\Lambda\ f+\delta)' \\ &= (\Lambda\ f+\delta)\ \left[(\Lambda\ f)'+\delta'\right] \\ &= (\Lambda\ f+\delta)\ (f'\Lambda'+\delta') \\ &= \Lambda\ f\ f'\Lambda'+\Lambda\ f\ \delta'+\delta\ f'\Lambda'+\delta\ \delta'\end{aligned}$$

A matriz de covariância ($\sum$) das variáveis y_i como desvios de suas médias é definida por:

$$\begin{aligned}\sum &= E\left[(y-\mu)(y-\mu)'\right] \\ &= E\left[(\Lambda\ f\ \ f'\Lambda'+\Lambda\ f\ \delta'+\delta\ \ f'\Lambda'+\delta\ \delta'\right] \\ &= \Lambda\ E(f\ \ f')\Lambda'+\Lambda E(f\ \ \delta')+E(\delta\ \ f')\Lambda'+E(\delta\ \delta') \\ &= \Lambda\ I\ \ \Lambda'+\Lambda\ O+O\ \ \Lambda'+\psi \\ &= \Lambda\ \Lambda'+\psi\end{aligned}$$

Assim tem-se:

$\sum = \text{COV(y)} = \Lambda\ \Lambda'+\psi$, ou seja, $\text{Var}(y_i) = \lambda_{i1}^2 + \lambda_{i1}^2 + ...\lambda_{im}^2 + \psi_i$

$\text{COV(y, f)} = \Lambda$, ou seja, $\text{COV}(y_i, f_j) = \lambda_{ij}$: covariância entre a i-ésima variável y_i com o fator comum f_j.

Define-se também os seguintes termos:

Comunalidade (c_i^2): porção da variância Var(y_i) da i-ésima variável y_i distribuída pelos m fatores comuns: $c_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + ...\lambda_{im}^2$: soma dos quadrados dos pesos da i-ésima variável sobre os fatores comuns.

Especificidade ou Variância Específica: porção da variância Var(y_i) devida ao fator específico: ψ_i.

Assim Var(y_i) pode ser decomposta em:

$$Var(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + ...\lambda_{im}^2 + \psi_i$$
$$= c_i^2 + \psi_i$$

= Comunalidade + variância específica.

Desta decomposição, tem-se que a comunalidade da variável i é dada também por $c_i^2 = Var(y_i) - \psi_i$.

O modelo fatorial assume que as p(p+1)/2 variâncias e covariâncias para y podem ser reproduzidas a partir dos pm pesos λ_{ij} e p variâncias ψ_i. Quando m = p, qualquer matriz de covariância $\sum$ pode ser reproduzida exatamente como $\Lambda\Lambda'$e, nesse caso, ψ será uma matriz nula. Com m < p, o modelo fatorial propicia uma interpretação mais simples da estrutura de covariância de y, usando menos parâmetros que os p(p+1)/2 elementos de $\sum$. Por exemplo, com p = 10 variáveis, o modelo fatorial com m = 2 fatores comuns necessita de pm + p = 10 x 2 + 10 = 30 parâmetros para explicar a estrutura da covariância associada aos p(p+1)/2 = 55 elementos.

Em aplicações práticas, a análise de fatores é geralmente realizada a partir de dados padronizados. Neste caso, a matriz de covariância $\sum$ é equivalente à matriz de correlação R dos dados originais. Assim, sob padronização, tem-se a igualdade $\sum$ = R = $\Lambda\Lambda' + \psi$. Também λ_{ij} agora representa o coeficiente de correlação entre a i-ésima variável e o j-ésimo fator comum. A diagonal da matriz C = R - ψ = $\Lambda\Lambda'$ fornece as comunalidades das variáveis. A comunalidade da variável i, no caso, é dada por $c_i^2 = 1 - \psi_i$ e é um indicativo da eficiência da representação da variável pelos fatores comuns. Valores de comunalidade superiores a 0,64 são adequados, pois refletem coeficientes de correlação superiores a 0,80 entre a variável e os fatores comuns (Souza, 1988). Segundo Cruz e Carneiro (2003), a qualidade média de toda a análise de fatores pode ser avaliada pela comunalidade média de todas as variáveis.

6.1.2 Estimação dos Carregamentos e Especificidades

Uma análise de fatores completa demanda a estimação dos carregamentos, especificidades e escores fatoriais. Na estimação dos referidos parâmetros, a estimação dos carregamentos é essencial e deve ser realizada primeiro, pois as estimativas de especificidades e dos escores fatoriais são função das estimativas dos elementos da matriz de carregamentos Λ. O peso ou carregamento de cada fator em cada variável refere-se à correlação entre o fator e a variável e os principais métodos para estimação de Λ são o dos componentes principais e o de máxima verossimilhança. A seguir, será apresentado o método dos componentes principais. Este método baseia-se na decomposição espectral de $\sum$ ou R tendo pares autovalor e autovetor (α_i, v_i) com $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_p \geq 0$.

Assim:

$$\sum = \alpha_1 v_1 v_1' + \alpha_2 v_2 v_2' + \cdots + \alpha_p v_p v_p'$$

$$= \left[(\alpha_1)^{1/2} v_1 \quad (\alpha_2)^{1/2} v_2 \quad \cdots \quad (\alpha_p)^{1/2} v_p \right] \begin{bmatrix} (\alpha_1)^{1/2} v_1' \\ (\alpha_2)^{1/2} v_2' \\ \vdots \\ (\alpha_p)^{1/2} v_p' \end{bmatrix}$$

Nesta decomposição foram consideradas as p variáveis, de forma que m = p e as variâncias específicas são nulas, isto é, $\psi_i = 0 \ \forall_i$. Portanto $\underset{pxp}{\sum} = \underset{pxp}{\Lambda} \ \underset{pxp}{\Lambda'} + \underset{pxp}{0} = \Lambda\Lambda'$. Esta decomposição é exata, mas não é útil.

Uma decomposição útil deve ter m pequeno. Quando os últimos p – m autovalores são pequenos, a contribuição de $\alpha_{m+1} v_{m+1} + \cdots + \alpha_p v_p v_p'$ é negligenciável em $\sum$ decomposta anteriormente. Desconsiderando esta contribuição, obtém-se:

$$\underset{pxp}{\sum} = \begin{bmatrix} (\alpha_1)^{1/2} v_1 & (\alpha_2)^{1/2} v_2 \cdots (\alpha_m)^{1/2} v_m \end{bmatrix} \begin{bmatrix} (\alpha_1)^{1/2} v_1' \\ (\alpha_2)^{1/2} v_2' \\ \vdots \\ (\alpha_m)^{1/2} v_m' \end{bmatrix}$$

$$\underset{pxp}{\sum} = \underset{pxm}{\Lambda} \; \underset{mxp}{\Lambda'}$$

Esta representação ignora os fatores específicos. Incluindo os fatores específicos no modelo, suas variâncias devem ser consideradas como os elementos diagonais de $\sum - \Lambda \Lambda'$, ou seja, ψ_i = diag ($\sum - \Lambda \Lambda'$).

Incluindo os fatores específicos, a decomposição espectral torna-se:

$$\sum = \Lambda \Lambda' + \psi = \begin{bmatrix} (\alpha_1)^{1/2} v_1 & (\alpha_2)^{1/2} v_2 \cdots (\alpha_m)^{1/2} v_m \end{bmatrix} \begin{bmatrix} (\alpha_1)^{1/2} v_1' \\ (\alpha_2)^{1/2} v_2' \\ \vdots \\ (\alpha_m)^{1/2} v_m' \end{bmatrix} + \psi$$

Em notação matricial tem-se:

$\sum = V D_\alpha V' + \psi$, em que:

$$V = \begin{array}{c} \begin{matrix} v_1 & v_2 & \cdots & v_m \end{matrix} \\ \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pm} \end{bmatrix} \end{array} = \text{matriz dos m autovetores}$$

$$D_\alpha = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \alpha_m \end{bmatrix} = \text{matriz diagonal dos m autovalores}$$

Verifica-se então a igualdade $\Lambda\ \Lambda' = VD_\alpha\ V'$, a qual pode ser expressa como $\Lambda\ \Lambda' = VD_\alpha^{1/2}\ D_\alpha^{1/2}\ V'$. Em conseqüência, nota-se que $\Lambda = VD_\alpha^{1/2}$.

Dessa forma, para determinação dos pesos ou carregamentos fatoriais, são necessários apenas os autovalores e autovetores das matrizes de covariância $\sum$ ou de correlação R. E isto é um problema puramente matemático de álgebra linear e de matrizes.

A análise de fatores expressa as variáveis como função dos fatores, ou seja, as variáveis são, em parte, explicadas pelos fatores. Por outro lado, na análise de componentes principais, os componentes são expressos como função das variáveis, ou seja, as variáveis explicam os componentes principais. O modelo para os componentes principais é dado por $CP_j = v_j y = v_{1j}\ y_1 + v_{2j}\ y_2 + ... + v_{pj}\ y_p$. Verifica-se, assim, que os elementos v_{ij} dos autovetores oriundos da decomposição espectral da matriz R são os próprios elementos ponderadores das variáveis nos componentes principais. Por outro lado, os pesos ou carregamentos ou ponderadores dos fatores na análise de fatores são dados por $\lambda_{ij} = v_{ij}(\alpha_j)^{1/2}$. Isto mostra a relação existente entre as técnicas de componentes principais e de análise de fatores. O modelo fatorial contém, adicionalmente, os fatores específicos.

O número de fatores a serem selecionados pode ser determinado por: (i) estrutura física das variáveis que pode sugerir um certo número de fatores; (ii) número de autovalores maiores que 1, quando se usa a matriz de correlação (critério de Kaiser, 1958); (iii) número de fatores que explicam uma proporção desejada (usualmente 80 %) da variação total.

6.1.3 Aplicação Prática

A seguir, considera-se o exemplo apresentado por Johnson e Wichern (1982) referente a cinco atributos de um alimento: y_1 = sabor; y_2 = preço; y_3 = aroma; y_4 = refeição ligeira; y_5 = nutrição.

A matriz de correlação R equivale a:

$$\begin{array}{c} \begin{matrix} y_1 & y_2 & y_3 & y_4 & y_5 \end{matrix} \\ \begin{bmatrix} 1,00 & 0,02 & 0,96 & 0,42 & 0,01 \\ & 1,00 & 0,13 & 0,71 & 0,85 \\ & & 1,00 & 0,50 & 0,11 \\ & & & 1,00 & 0,79 \\ & & & & 1,00 \end{bmatrix} \end{array} \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{matrix}$$

Os autovalores da matriz R são:

$$\hat{\alpha}_1 = 2,85; \quad \hat{\alpha}_2 = 1,81; \quad \hat{\alpha}_3 = 0,20; \quad \hat{\alpha}_4 = 0,10; \quad \hat{\alpha}_5 = 0,03$$

Verifica-se a presença de dois autovalores maiores que 1. Portanto, m = 2 fatores é suficiente segundo o critério de Kaiser. Verifica-se também que os dois fatores explicam 93 % da variação total, ou seja, $(\hat{\alpha}_1 + \hat{\alpha}_2)/p = (2,85 + 1,81)/5 = 0,93$. Isto confirma que a adoção de m = 2 fatores é adequada. Também, observando-se a matriz R, verifica-se dois grupos de variáveis com correlações altas entre si: variáveis 1 e 3 formando um primeiro grupo ou fator e variáveis 2, 4 e 5 formando um segundo grupo ou fator.

Os autovetores associados a $\hat{\alpha}_1$ e $\hat{\alpha}_2$ são:

$$\hat{v}_1 = [0,33 \quad 0,46 \quad 0,38 \quad 0,56 \quad 0,47]$$

$$\hat{v}_2 = [0,61 \quad -0,39 \quad 0,56 \quad -0,07 \quad 0,40]$$

A matriz de carregamentos é estimada por:

$$\hat{\Lambda} = VD_{\alpha}^{1/2} = \begin{array}{c} \begin{matrix} \hat{\lambda}_{i1} & \hat{\lambda}_{i2} \end{matrix} \\ \begin{bmatrix} 0,33 & 0,61 \\ 0,46 & -0,39 \\ 0,38 & 0,56 \\ 0,56 & -0,07 \\ 0,47 & -0,40 \end{bmatrix} \end{array} \begin{bmatrix} (2,85)^{1/2} & 0 \\ 0 & (1,81)^{1/2} \end{bmatrix} = \begin{bmatrix} 0,56 & 0,82 \\ 0,78 & -0,52 \\ 0,64 & 0,75 \\ 0,95 & -0,09 \\ 0,79 & -0,54 \end{bmatrix}$$

em que $\hat{\lambda}_{i1}$ e $\hat{\lambda}_{i2}$ são os carregamentos dos fatores 1 e 2 na variável i, respectivamente.

As comunalidades (porção da variância da i-ésima variável compartilhada com os dois fatores) são estimadas por:

$$\hat{c}_1^2 = \hat{\lambda}_{i1}^2 + \hat{\lambda}_{i2}^2$$

$$\hat{c}_1^2 = \hat{\lambda}_{11}^2 + \hat{\lambda}_{12}^2 = 0{,}56^2 + 0{,}82^2 = 0{,}99$$
$$\hat{c}_2^2 = 0{,}78^2 + 0{,}52^2 = 0{,}88$$
$$\hat{c}_3^2 = 0{,}97$$
$$\hat{c}_4^2 = 0{,}91$$
$$\hat{c}_5^2 = 0{,}92$$

Sendo C a matriz das comunalidades, as variâncias específicas são estimadas por:

$$\hat{\psi} = diag(R - \Lambda\Lambda') = diag(R - C) = diag(I - C)$$
$$\psi_i = 1 - \hat{c}_i^2$$
$$\psi_1 = 1 - 0{,}99 = 0{,}01$$
$$\psi_2 = 1 - 0{,}88 = 0{,}12$$
$$\psi_3 = 0{,}03$$
$$\psi_4 = 0{,}09$$
$$\psi_5 = 0{,}008$$

O resumo dos resultados encontra-se na Tabela a seguir:

Variável	**Carregamentos** $(\hat{\lambda}_{ij})$		**Comunalidades** $(\hat{c}_i^2)$	**Variância Específica** $(\hat{\psi}_i)$
	f_1	**f_2**		
1-Sabor	0,56	0,82	0,99	0,01
2-Preço	0,78	-0,52	0,88	0,12
3-Aroma	0,64	0,75	0,97	0,03
4-Refeição Ligeira	0,95	-0,09	0,91	0,09
5-Energia	0,79	-0.54	0,92	0,08
Autovalores	$\hat{\lambda}_1 = 2{,}85$	$\hat{\lambda}_2 = 1{,}81$		
Proporção Acumulada	**57%**	**93%**		

A proporção acumulada pode ser dada também por :

$$\sum \hat{c}_i^2 / p = (0{,}99 + 0{,}88 + 0{,}97 + 0{,}91 + 0{,}92) / 5 = 0{,}93$$

A eficiência do modelo fatorial com m = 2 pode ser verificada também pela reconstituição da matriz R, a partir de $\hat{R} = \hat{\Lambda}\hat{\Lambda}' + \hat{\psi}$. Tem-se:

$$\hat{R} = \hat{\Lambda}\hat{\Lambda}' + \hat{\psi} = \begin{bmatrix} 0{,}56 & 0{,}82 \\ 0{,}78 & -0{,}52 \\ 0{,}64 & 0{,}75 \\ 0{,}95 & -0{,}09 \\ 0{,}79 & -0{,}54 \end{bmatrix} \begin{bmatrix} 0{,}56 & 0{,}78 & 0{,}64 & 0{,}95 & 0{,}79 \\ 0{,}82 & -0{,}52 & 0{,}75 & -0{,}09 & -0{,}54 \end{bmatrix} + \begin{bmatrix} 0{,}01 & & & & 0 \\ & 0{,}12 & & & \\ & & 0{,}03 & & \\ & & & 0{,}09 & \\ 0 & & & & 0{,}08 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0{,}01 & 0{,}97 & 0{,}46 & 0{,}00 \\ & 1 & 0{,}11 & 0{,}79 & 0{,}90 \\ & & 1 & 0{,}59 & 0{,}10 \\ & & & 1 & 0{,}80 \\ & & & & 1 \end{bmatrix}$$

Verifica-se que a matriz de correlação estimada pelo modelo fatorial aproxima bem a matriz de correlação inicial. Isto demonstra a eficiência do modelo com m = 2 fatores. As altas comunalidades indicam que os dois fatores explicam uma grande porcentagem de variação de cada variável.

6.1.4 Rotação dos Fatores

Na situação em que os carregamentos estimados não fornecem uma interpretação fácil da ênfase dada aos fatores nas variáveis, é ideal rotacioná-los na busca de uma estrutura com interpretação mais simples. O ideal é a obtenção de uma estrutura para os carregamentos tal que cada variável tenha peso alto em um único fator e pesos baixos ou moderados nos demais fatores.

Da álgebra matricial emerge que uma transformação ortogonal corresponde a uma rotação rígida dos eixos coordenados. Em análise fatorial, uma transformação ortogonal é denominada rotação de fatores ou rotação fatorial. Sendo $\hat{\Lambda}$ a matriz pxm dos carregamentos estimados, $\hat{\Lambda}^* = \hat{\Lambda}T$ (onde TT´=T´T=I) é uma matriz pxm dos carregamentos rotacionados. A matriz de correlação (ou de

covariância) estimada permanece inalterada, pois: $\hat{\Lambda}\hat{\Lambda}' + \hat{\psi} = \hat{\Lambda}TT'\hat{\Lambda}' + \hat{\psi} = \hat{\Lambda}T(\hat{\Lambda}'T)' + \hat{\psi} = \hat{\Lambda}^* \hat{\Lambda}^{*\prime} + \psi$, em que T é uma matriz ortogonal. As comunalidades e variâncias específicas também permanecem inalteradas com a rotação. Assim, do ponto de vista matemático, é indiferente obter $\hat{\Lambda}$ ou $\hat{\Lambda}^*$.

Quando m = 2, a transformação para uma estrutura simples pode ser determinada graficamente. Toma-se f_1 e f_2 como eixos coordenados ortogonais (f_1 e f_2 são independentes e não correlacionados) e plota-se os pares de pesos $(\hat{\lambda}_{i1},\ \hat{\lambda}_{i2})$ para os p pontos ou variáveis. Os eixos coordenados podem então ser rotacionados de um ângulo ϕ e os novos pesos rotacionados $\hat{\lambda}^*_{ij}$ são determinados pela relação $\underset{px2}{\hat{\Lambda}^*} = \underset{px2}{\hat{\Lambda}}\ \underset{2x2}{T}$, em que:

$$T = \begin{bmatrix} \cos\phi & \text{sen}\,\phi \\ -\text{sen}\,\phi & \cos\phi \end{bmatrix}$$ para uma rotação no sentido horário.

Para m>2, esta análise é impraticável e programas computacionais são utilizados para tal. Outro critério para obter uma estrutura mais simples é conhecido como **rotação VARIMAX** (Kaiser, 1958), a qual é baseada em alogarítmos computacionais.

Para o exemplo do item anterior, a rotação fatorial, considerando $\phi = 33{,}3^o$, é dada por:

$$\hat{\Lambda}^* = \hat{\Lambda}T = \begin{bmatrix} 0{,}56 & 0{,}82 \\ 0{,}78 & -0{,}52 \\ 0{,}64 & 0{,}75 \\ 0{,}95 & -0{,}09 \\ 0{,}79 & -0{,}54 \end{bmatrix} \begin{bmatrix} \cos\ 33{,}3^o & \text{sen}\ 33{,}3^o \\ -\text{sen}\ 33{,}3^o & \cos\ 33{,}3^o \end{bmatrix} = \begin{bmatrix} 0{,}02 & 0{,}99 \\ 0{,}94 & -0{,}01 \\ 0{,}13 & 0{,}98 \\ 0{,}84 & 0{,}43 \\ 0{,}97 & -0{,}02 \end{bmatrix} = \hat{\Lambda}^*$$

Verifica-se que a rotação fatorial tornou mais clara a ênfase dada aos fatores por cada variável. Verifica-se que as variáveis 2, 4 e 5 apresentam pesos mais altos no fator 1 e pesos relativamente mais baixos no fator 2. As variáveis 1 e 3 apresentam pesos mais altos no fator 2 e pesos mais baixos no fator 1. Assim, os dois fatores foram claramente definidos e as variáveis puderam ser agrupadas em dois grupos, de acordo com suas correlações.

6.1.5 Escores Fatoriais

Além das estimativas dos parâmetros (carregamentos e especificidades) do modelo fatorial, a estimação dos próprios fatores comuns, denominados escores fatoriais, são úteis em várias aplicações tais quais análise de agrupamento. Nos escores fatoriais, os fatores comuns são expressos como combinações lineares das características. Assim, são obtidos escores fatoriais para cada indivíduo ou objeto.

Em termos da matriz total dos dados associada a n indivíduos e p variáveis, ou seja, de uma matriz n x p de dados, tem-se o seguinte sistema linear:

$Y = F\Lambda + \Delta$, em que:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} =$$ matriz de observações de p variáveis em n indivíduos

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{bmatrix} =$$ matriz dos m fatores comuns associados aos n indivíduos ou matriz dos escores fatoriais dos n indivíduos, em que cada linha contém os escores fatoriais associados a cada indivíduo.

Λ= matriz de carregamentos fatoriais, já definida anteriormente.

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1p} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2p} \\ \vdots & \vdots & & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{np} \end{bmatrix} =$$ matriz dos p fatores específicos associados aos n indivíduos.

Os escores fatoriais obtidos pelo método de quadrados mínimos ordinários a partir das estimativas dos carregamentos pelo método dos componentes principais são dados por:

$$\hat{F} = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'Y$$

Assim, os ponderadores dos valores observados das variáveis para obtenção dos escores fatoriais são dados por $\hat{p} = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'$, ou seja, dependem apenas dos carregamentos. Os coeficientes de ponderação associados ao exemplo anterior são:

$$\hat{p} = \begin{bmatrix} 0{,}19 & 0{,}27 & 0{,}22 & 0{,}33 & 0{,}28 \\ 0{,}45 & -0{,}29 & 0{,}41 & -0{,}05 & -0{,}30 \end{bmatrix}$$

Dessa forma, os escores fatoriais para cada indivíduo dados por:

$$\hat{f}_1 = 0{,}19\ y_1 + 0{,}27\ y_2 + 0{,}22\ y_3 + 0{,}33\ y_4 + 0{,}28\ y_5$$
$$\hat{f}_2 = 0{,}45\ y_1 + 0{,}29\ y_2 + 0{,}41\ y_3 + 0{,}05\ y_4 + 0{,}30\ y_5$$

Substituindo-se os valores de y_1 .. y_5 de cada indivíduo, obtém-se duas novas variáveis para todos os indivíduos. Estas novas variáveis são denominadas supervariáveis. O número de supervariáveis será igual ao número m de fatores e podem ser sujeitas a novas análises estatísticas.

Os escores fatoriais podem também ser estimados em associação com os modelos FAMM, conforme apresentado no Capítulo 8.

6.1.6 Uso da Análise de Fatores no Melhoramento Genético

A análise de fatores pode ser aplicada ao melhoramento genético, principalmente quanto aos objetivos relatados a seguir.

a) Agrupamento de caracteres segundo suas correlações (conforme no exemplo apresentado)

b) Simplificação e consideração da estrutura de correlação entre caracteres no estabelecimento de índices de seleção multicaracterísticos

O índice de seleção tradicional de Smith-Hazel considera, simultaneamente, todos os caracteres de interesse, gerando uma variável adicional (que é um componente principal) que resulta da

ponderação dos caracteres por meio de coeficientes calculados com base nas herdabilidades, valores econômicos relativos e correlações genéticas e fenotípicas entre os caracteres. Os componentes principais e os escores fatoriais são também índices de seleção multicaracterísticos, os quais consideram apenas as correlações (genéticas ou fenotípicas) entre os caracteres. Não consideram os valores econômicos relativos e nem as herdabilidades dos caracteres. Este último fato justifica a realização de novas análises estatístico-genéticas sobre as supervariáveis geradas. Outra diferença entre o índice de seleção tradicional e os escores fatoriais e dos componentes principais refere-se ao fato de que nestes últimos são gerados mais de um eixo principal, associados aos fatores e componentes principais. Em outras palavras, com m = 2, são geradas duas variáveis adicionais para cada indivíduo.

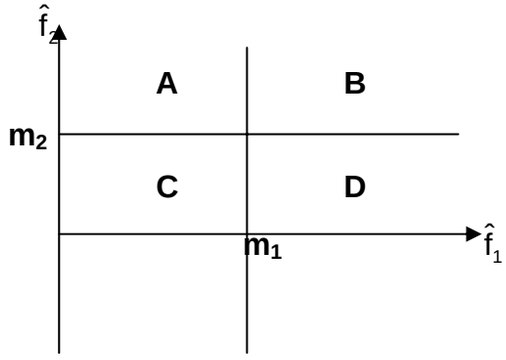
A eficiência da seleção baseada nestes três tipos de índice vai depender, sobretudo, da estrutura de correlação observada e das herdabilidades dos caracteres. Em termos conceituais, o índice tradicional tende a ser mais completo. Os escores fatoriais e componentes principais têm a vantagem de usar uma estrutura de correlação simplificada e parcimoniosa, fato que pode conduzir a uma maior eficiência de estimação. Os três tipos tendem a ser superiores à seleção baseada em caracteres individuais. Estes tipos de seleção multivariada podem ser superiores à seleção univariada em situações que um caráter funciona como auxiliar na seleção do outro. E isto ocorre, sobretudo, quando a diferença entre as correlações genética e ambiental entre dois caracteres é de elevada magnitude (Resende, 2002, p.331).

c) Análise da interação genótipo x ambiente, visando a estratificação de ambientes de plantio

As variáveis referidas na análise de fatores podem ser representadas por diferentes locais de experimentação. Neste caso, os locais podem ser agrupados em fatores de acordo com o carregamento nos fatores, os quais refletem as correlações entre os locais. Os carregamentos rotacionados do primeiro e segundo fatores para cada local podem ser plotados em um sistema de coordenadas, o qual exibirá o agrupamento dos locais.

d) Análise de adaptabilidade dos materiais genéticos

Com m = 2 fatores selecionados, os escores fatoriais médios dos genótipos em cada fator podem ser utilizados para divisão de um quadrante do sistema cartesiano (com f_1 na abscissa X e f_2 na ordenada Y) em quatro sub-quadrantes da seguinte forma (Cruz e Carneiro, 2003):



em que:

m_1 = média dos escores fatoriais dos genótipos no fator 1;

m_2 = média dos escores fatoriais dos genótipos no fator 2;

A = quadrante ocupado por genótipos adaptados aos locais do grupo 2 (indicado pelo fator 2), ou seja, com adaptabilidade específica ao grupo 2;

B = quadrante ocupado por genótipos com adaptabilidade ampla, ou seja, com adaptação aos dois grupos de locais;

C = quadrante ocupado por genótipos não selecionáveis, ou seja, sem adaptabilidade alguma;

D = quadrante ocupado por genótipos com adaptabilidade específica aos locais do grupo 1 (indicado pelo fator 1).

Plotando-se no gráfico acima os pontos $(\hat{f}_1;\ \hat{f}_2)$ associados aos vários genótipos, haverá automaticamente a alocação dos mesmos aos diferentes quadrantes e, portanto, a classificação dos mesmos em termos de adaptabilidade. Cruz e Carneiro (2003) discutem a aplicação da análise de

fatores no estudo da adaptabilidade. Piepho (1998), Smith, Cullis e Thompson (2001) e Resende e Thompson (2003) também usam a análise de fatores com este objetivo, porém no contexto dos modelos mistos (REML/BLUP), conforme apresentado no Capítulo 8.

e) Predição de valores genéticos, estratificação ambiental e classificação dos genótipos por adaptabilidade e estabilidade via REML/BLUP

Esta é uma análise completa, adequada a dados desbalanceados, a situação de heterogeneidade de variância entre locais e a presença de variabilidade espacial dentro de ensaios. Adicionalmente, os efeitos de genótipos e de interação g x e são tratados como aleatórios, fato que conduz a melhores predições. Este procedimento é apresentado em detalhes no Capítulo 8.

6.2 Análise de Fatores sob Modelos Multiplicativos Mistos (FAMM)

A estrutura da matriz de covariância ou correlação envolvendo v caracteres está associada a v(v+1)/2 elementos. Visando simplificar a estrutura dessa matriz, sumarizar a informação multivariada e reduzir a dimensionalidade do problema, decomposições dessas matrizes, baseadas em seus autovalores e autovetores, são usadas com base em diferentes parametrizações produzindo as técnicas de componentes principais e da análise de fatores.

Entretanto, tais procedimentos são baseados na estimação completa da matriz de covariância ou de correlação com todos os seus v(v+1)/2 elementos. Um procedimento estatístico mais atrativo refere-se a estimar os componentes principais e os fatores diretamente, restringindo a estimação apenas àqueles mais importantes. Esse procedimento não requer a estimação prévia da matriz de covariância e de correlação e é sobretudo relevante no contexto dos modelos mistos e de dados desbalanceados. Nesse caso, torna-se necessária uma reparametrização dos modelos mistos tradicionais (Smith et al., 2001; Resende e Thompson, 2003 e 2004; Meyer e Kirkpatrick, 2005).

A seguir é apresentada uma extensão dos modelos mistos para incorporar a análise de fatores.

Modelo misto tradicional

$y = Xb + Za + e$

Modelo misto fator analítico (FAMM)

$y = Xb + Z[(\Lambda \otimes I_g) f + \delta] + \varepsilon$, em que:

$a = [(\Lambda \otimes I_g) f + \delta]$

Sob esse modelo, a matriz de covariância genética é dada por $\sum = \Lambda\ \Lambda' + \psi$, em que $\Lambda\ \Lambda'$ = $VD_\alpha\ V'$, D_α é a matriz diagonal dos m autovalores e V é a matriz dos autovetores. Escolhendo-se V e D_α referentes apenas à dimensão m, esse modelo misto é reduzido e ajusta somente os m fatores. Na técnica FAMM, a estrutura de covariância é simplificada para $\sum = \Lambda_m \Lambda_m + \Psi'$.

A metodologia de modelos mistos padrão pode ser usada para estimar autovalores e autovetores diretamente sem a necessidade de se estimar Σ completa. A principal diferença para o modelo multivariado misto tradicional refere-se ao fato de que os parâmetros a serem estimados fazem parte da matriz de incidência dos efeitos genéticos aleatórios. Como a distribuição de $[(\Lambda \otimes I_g) f$ é singular, isto conduz à estimação sob posto reduzido, restrições devem ser impostas aos parâmetros do modelo fator analítico. Thompson et al. (2003) descrevem o procedimento adequado para a estimação de parâmetros nessa situação. Detalhes do ajuste de modelos mistos fator analíticos são apresentados no Capítulo 8.

Uma importante aplicação dos modelos fator analíticos mistos é na análise de múltiplos experimentos e estudos de interação genótipo x ambiente. Esta é provavelmente a melhor forma de predição de valores genéticos e determinação da estabilidade e adaptabilidade de genótipos através dos ambientes, superando métodos tais quais o AMMI e outros métodos tradicionais. A técnica FAMM reúne em um só método os procedimentos de análise multivariada, análise de adaptabilidade e estabilidade e análise de modelos mistos.

A associação das técnicas de análise multivariada e de modelos mistos é importante para a análise de múltiplos experimentos e também múltiplos caracteres e, em alguns casos, medidas repetidas. Para o caso de múltiplos caracteres, o uso da PCAM é mais adequado. Para múltiplos

experimentos, a técnica FAMM é mais indicada. Isto porque a análise de componentes principais enfatiza a identificação de variáveis que explicam o máximo da variação total, fato que é relevante para o caso de múltiplos caracteres. Por outro lado, a análise de fatores enfatiza a atribuição da covariância entre variáveis a fatores comuns. Isto é relevante quando as variáveis referem-se a ambientes ou experimentos e todos os ambientes são alvo da análise e não apenas aqueles que mais contribuem para a variação total. Também, a covariância ou correlação entre ambientes atribuídas a fatores comuns automaticamente considera a similaridade e dissimilaridade entre ambientes, o que é uma propriedade interessante nesse contexto. Uma descrição detalhada e exemplo de aplicação da técnica FAMM na análise de múltiplos experimentos com interação g x e é apresentada no Capítulo 8.

CAPÍTULO 8
ANÁLISE DE MÚLTIPLOS EXPERIMENTOS, ESTABILIDADE E ADAPTABILIDADE

1 INTERAÇÃO GENÓTIPOS X AMBIENTES

1.1 Conceitos e Implicações da Interação Genótipos X Ambientes

A interação genótipos x ambientes é definida como o comportamento diferencial de genótipos em diferentes ambientes. Na análise de um grupo de indivíduos avaliados em um ambiente ou local, o modelo para o valor fenotípico é: $y = \mu + g + e$.

Quando grupos de indivíduos são avaliados em mais de um local, o modelo para o valor fenotípico deve ser expandido para:

$y = \mu^* + \ell + g^* + g^*\ell + e$, em que:

μ^* : efeito da média geral livre do efeito de local.

g^* : efeito genotípico, livre da interação genótipo x ambiente.

ℓ: é o efeito de local ou macroambiente.

$g^*\ell$: é o efeito da interação genótipo x ambiente.

e: efeito do erro aleatório ou do microambiente.

Comparando este modelo com o anterior, têm-se as igualdades: $\mu = \mu^* + \ell$; $g = g^* + g^*\ell$. Assim, verifica-se que quando a avaliação é realizada em um único ambiente, a média geral (μ) é inflacionada pelo efeito de locais e o efeito genotípico (g), pelo efeito da interação genótipo x ambiente. O efeito da interação genótipos x ambientes é decorrente do comportamento diferencial dos diferentes genótipos nos diferentes ambientes e pode indicar que os melhores indivíduos em um ambiente podem não sê-lo em outro. Assim, este pode ser um complicador na seleção, se não for considerado adequadamente.

A variação fenotípica total é dada por $\sigma^2_y = \sigma^2_{g^*} + \sigma^2_{g^*l} + \sigma^2_e$ (considerando fixo o efeito de locais).

No caso de dois ambientes, as herdabilidades relativas às várias modalidades de seleção são dadas conforme a Tabela 24, obtida com base em Resende (1989).

Tabela 24. Herdabilidades associadas à seleção em dois ambientes.

Ambiente de seleção	Ambiente onde se deseja o ganho	Herdabilidade*
X	X	$h^2_{x/x} = \frac{\sigma^2_{g_x}}{\sigma^2_x}$
X	Y	$h^2_{y/x} = \frac{\sigma^2_{g^*}}{\sigma^2_x}$
X	Média dos ambientes x e y	$h^2_{(x,y)/x} = \frac{(1/2)(\sigma^2_{g^*} + \sigma^2_{g_x})}{\sigma^2_x}$
Média dos ambientes x e y	X	$h^2_{x/(x,y)} = \frac{(1/2)(\sigma^2_{g^*} + \sigma^2_{g_x})}{\sigma^2_{(x,y)}}$
Média dos ambientes x e y	Média dos ambientes x e y	$h^2_{((x,y)/(x,y)} = \frac{(1/4)\ (\sigma^2_{gx} + \sigma^2_{gy} + 2\sigma^2_{g^*})}{\sigma^2_{(x,y)}} = \frac{\sigma^2_{g^*} + (1/2)\ \sigma^2_{g^*\ell}}{\sigma^2_{(x,y)}}$

* σ^2_x e $\sigma^2_{(x,y)}$: variâncias fenotípicas no ambiente x e na média dos ambientes x e y, respectivamente.

σ^2_{gx} : variância genética no local x.

$\sigma^2_{g^*\ell}$: variância da interação genótipos com ambientes.

$\sigma^2_{g^*} + \sigma^2_{g^*\ell} = (1/2)(\sigma^2_{gx} + \sigma^2_{gy})$.

Verifica-se que quando um material é avaliado em um ambiente e utilizado no mesmo, o efeito da interação é capitalizado na seleção, ou seja, o valor genotípico dos indivíduos selecionados e utilizados no local ℓ é dado por: $\mu^* + \ell + g^* + g^*\ell$.

Quando o material é avaliado em um ambiente e utilizado em outro, o valor genotípico no outro ambiente é menor, equivalendo $\mu^* + g^*$. Assim, a interação pode ser favorável ou desfavorável à seleção, dependendo da estratégia de utilização do material selecionado.

1.2 Correlação Genética Através dos Ambientes e Número de Locais de Experimentação

O componente de variância da interação genótipo x ambiente pode ser assim desdobrado (Robertson, 1959):

$\sigma^2_{g^*\ell} = \frac{1}{2}(\sigma_{gx} - \sigma_{gy})^2 + (1 - r_{gxgy})\ \sigma_{gx}\ \sigma_{gy}$, em que:

$\frac{1}{2}(\sigma_{gx} - \sigma_{gy})^2$: parte simples da interação, explicada pela mudança de variação genética (heterogeneidade de variâncias) dos materiais nos diferentes ambientes.

$(1 - r_{gxgy})\ \sigma_{gx}\ \sigma_{gy}$: parte complexa da interação, advinda da falta de correlação genética entre o desempenho do material genético de um ambiente para outro. É a parte problemática da interação, significando que o material bom em um ambiente pode não sê-lo em outro, quando r_{gxgy} é baixa.

r_{gxgy}: correlação genética entre o desempenho do material genético de um ambiente para outro.

σ_{gx}: desvio padrão genético aditivo no ambiente x.

Pela expressão de $\sigma^2_{g^*\ell}$, verifica-se que, quando a correlação genética através dos ambientes equivale a 1, a variância da interação genótipo x ambiente refere-se, exclusivamente, à heterogeneidade de variâncias.

A correlação r_{gxgy} pode ser estimada pelas expressões:

$$\hat{r}_{gxgy} = \frac{\sigma^2_{g*}}{\sigma_{gx}\ \sigma_{gy}} \text{ ou}$$

$$\hat{r}_{gxgy} = \frac{\sigma^2_{g*}}{\sigma^2_{g*} + \sigma^2_{g*\ell} - 0{,}5\ (\sigma_{gx} - \sigma_{gy})^2} \text{ ou}$$

$$\hat{r}_{gxgy} = r_{xy}/(h_x\ h_y),$$

em que r_{xy} é a correlação fenotípica entre o desempenho do material genético de um ambiente para o outro e h_x e h_y são as raízes quadradas das herdabilidades nos ambientes x e y, respectivamente. Esta expressão é válida, pois a covariância fenotípica (numerador da correlação fenotípica), neste caso, é somente de natureza genética.

Esta correlação $r_{gx\ gy}$ é também denominada correlação genética do tipo B, por ser avaliada sobre diferentes indivíduos.

Para situações que envolvem mais que dois locais, a correlação genética pode ser estimada por:

$$\hat{r}_{gg} = \frac{\sigma^2_{g*}}{\sigma^2_{g*} + \sigma^2_{g*\ell} - V(\sigma_{g_i})}, \text{ em que:}$$

$V(\sigma_{g_i})$: variância da escala genética ou variância dos desvios padrões genéticos nos ambientes. Em caso de variações genéticas de magnitudes similares nos ambientes, este valor torna-se nulo.

Na estimação de $r_{gg}\ e\ r_{gxgy}$, o conceito do coeficiente de correlação genotípica intraclasse pode ser também utilizado (Dickerson, 1962). Considerando homogeneidade de variância genética ($V(\sigma_{g_i}) = 0$) nos ambientes, tem-se que:

$$\hat{r}_{gxgy} = \frac{\sigma^2_{g*}}{\sigma_{gx}\ \sigma_{gy}} = \frac{\sigma^2_{g*}}{\sigma^2_{g*} + \sigma^2_{g*\ell}}$$

Este estimador subestima a correlação genética se houver heterogeneidade de variâncias e, neste caso, é um estimador conservador do ponto de vista do melhorista, já que superestimará (diminuindo a estimativa da correlação genética) o efeito da interação e, conseqüentemente, levará o melhorista a se precaver contra ela. Assim, r_{gg} (assumindo $V(\sigma_{gi}) = 0$) é o limite inferior da correlação genética esperada para um caráter avaliado em parentes em dois diferentes ambientes. Equivale também a um estimador exato de r_{gxgy} quando se corrige (padronização), previamente, a heterogeneidade de variância dos dados. Este estimador pode ser aplicado facilmente a partir de uma análise de variância conjunta, segundo um modelo aleatório, como no esquema mostrado a seguir.

Fontes de variação	QM	E(QM)	Teste F
Ambientes (L)	-	-	-
Genótipos (G)	Q_1	$\sigma^2 + n\ \sigma^2_{g*\ell} + n\ell\ \sigma^2_{g*}$	$F = Q_1 / Q_2$
G x L	Q_2	$\sigma^2 + n\ \sigma^2_{g*\ell}$	$F^* = Q_2 / Q_3$
Resíduo	Q_3	σ^2	-

n e ℓ : número médio de indivíduos por progênies (genótipos) e número de ambientes de avaliação, respectivamente

A partir desta análise, têm-se os seguintes estimadores:

$$\hat{\sigma}^2_{g*\ell} / \hat{\sigma}^2_{g*} = \frac{(1 - 1/F^*)}{(F-1)}\ell = \frac{(F^* - 1)}{(F-1)}\ \frac{\ell}{F^*}$$

$$\hat{r}_{gg} = \frac{\sigma^2_{g*}}{\sigma^2_{g*} + \sigma^2_{g*\ell}} = \frac{(F-1)}{(F-1) + (\ell - \ell / F^*)} = \frac{1}{1 + \frac{(F^* - 1)}{(F-1)}\ \frac{\ell}{F^*}}$$

As correlações $r_{gg}\ e\ r_{gxgy}$ podem ser expressas alternativamente em função da proporção $P = \sigma^2_{g*\ell} / \sigma^2_{g*}$, por meio de $r_{gg} = \sigma^2_{g*} / (\sigma^2_{g*} + P\ \sigma^2_{g*}) = 1/(1+P)$. Com P = 0,5, tem-se $r_{gg} = 0{,}67$, que é um valor alto de correlação genética. Dessa forma, pode-se inferir que quando a proporção variância da interação/variância genética livre da interação for inferior a 0,5, a interação não é problemática para o melhorista, pois conduzirá a uma alta correlação. Quando P > 0,5, a interação poderá ser problemática para o melhorista, implicando perdas de ganho com a seleção indireta (seleção em um local visando a ganho em outro). Tem-se também a igualdade $P = \sigma^2_{g*\ell} / \sigma^2_{g*} = \frac{1 - r_{gg}}{r_{gg}}$.

A eficiência (E_f) da seleção baseada na média de vários (ℓ) ambientes em relação à seleção em um só ambiente pode ser inferida (para herdabilidades ao nível de médias, similares e tendendo a 1 nos vários ambientes) pela expressão: $E_f = [\ell / [1 + (\ell - 1)\ r_{gg}]]^{1/2}$. Esta expressão é função de $\left[\frac{h^2_{(x,y)/(x,y)}}{h^2_{(x,y)/x}}\right]^{1/2}$, em que estas herdabilidades estão definidas na Tabela 24. A expressão de E_f mede a eficiência do uso de ℓ locais ao invés de um, visando a ganho na média dos ℓ locais.

Com base nesta expressão, foi estabelecida a Tabela 25.

Tabela 25. Eficiência (em termos de ganho genético na média dos locais) da utilização de ℓ locais de avaliação dos materiais genéticos em vez de apenas um, para diferentes valores da correlação genética (r_{gg}) envolvendo a performance do germoplasma nos ambientes.

r_{gg}	ℓ	E
0,95	2	1,01
	3	1,02
0,90	2	1,03
	3	1,04
0,85	2	1,04
	3	1,05
0,80	2	1,05
	3	1,07
0,75	2	1,07
	3	1,10
0,70	2	1,08
	3	1,12
0,65	2	1,10
	3	1,14
0,60	2	1,12
	3	1,17
0,55	2	1,14
	3	1,20
0,50	2	1,15
	3	1,22
0,40	2	1,20
	3	1,29
	4	1,35
0,30	2	1,24
	3	1,37
	4	1,45
	5	1,51

Verifica-se pela Tabela 25 que, quando a correlação genética for igual ou superior a 0,70, o ganho em eficiência, pela utilização de mais de um local de experimentação, é inferior a 10 %. Se a correlação genética for superior a 0,80, o ganho em eficiência é inferior a 5 %. Por outro lado, a utilização de três em vez de dois locais parece ser recomendada somente quando a correlação (estimada com base em três ou mais locais) for inferior a 0,5.

É relevante também determinar o número adequado de locais de experimentação, considerando um número total fixo de indivíduos a serem avaliados. Esse número depende da herdabilidade do caráter e da correlação genética intraclasse através dos locais. Fixando-se em $n\ell$ o número total de indivíduos por acesso, em que n refere-se ao número de indivíduos por local, considerando a avaliação dos $n\ell$ indivíduos alternativamente em um só ambiente ou em vários ambientes, a eficiência da seleção baseada em vários locais em relação à seleção baseada em um só local é dada por (Resende, 2002a):

$$E = \left[\frac{1 + (n\ell - 1)\ \hat{h}_i^2}{1 + (n-1)\ \hat{h}_i^2 + n(\ell - 1)\ \hat{r}_{gg}\ \hat{h}_i^2} \right]^{1/2}$$ em que $\hat{h}_i^2$ é estimativa da herdabilidade individual dentro de locais.

Na Tabela 26, são apresentados valores de eficiência para alguns valores de $n,\ \ell,\ e\ r_{gg}\ \ para\ \ h^2 = 0{,}20.$

Tabela 26. Eficiência (E) da experimentação em vários locais (l) em relação a um só local, considerando um número total (nl) fixo de indivíduos avaliados, para várias magnitudes de correlação genética (r_{gg}) intraclasse, através dos locais, número (n) de indivíduos por acesso em cada local e herdabilidade (h^2) de 20 %.

h^2	nl	l	n	r_{gg}	E	Incremento (%)
0,20	30	2	15,0	0,30	1,20	20
		3	10,0	0,30	1,30	10
		4	7,5	0,30	1,38	8
		5	6,0	0,30	1,40	2
0,20	30	2	15,0	0,50	1,13	13
		3	10,0	0,50	1,19	6
		4	7,5	0,50	1,22	3
		5	6,0	0,50	1,24	2
0,20	30	2	15,0	0,70	1,07	7
		3	10,0	0,70	1,10	3
		4	7,5	0,70	1,12	2
		5	6,0	0,70	1,13	1

Para um número total fixo de indivíduos avaliados, conclui-se, com base na Tabela 26, que torna-se vantajosa (ganho 6 % ou mais) a utilização de quatro, três ou dois locais para correlações de magnitudes da ordem de 0,30; 0,50 e 0,70, respectivamente. Este resultado mostra que a interação pode ser minimizada, mesmo sem o aporte de recursos adicionais, bastando dividir um grande experimento em vários locais.

2 VISÃO GERAL DOS MÉTODOS DE ANÁLISE DE MÚLTIPLOS EXPERIMENTOS, ESTABILIDADE E ADAPTABILIDADE

Os experimentos repetidos em vários ambientes são comuns na experimentação agrícola. As análises destes tipos de experimento objetivam a realização de inferências: (i) para ambientes

individuais; (ii) para o ambiente médio; (iii) para ambientes novos não incluídos na rede experimental. De maneira geral, a capacidade de os materiais genéticos se comportarem bem em uma grande amplitude de condições ambientais pode ser fator essencial em um programa de melhoramento genético. Neste contexto, o estudo da estabilidade dos materiais genéticos torna-se relevante. De maneira geral, um material é considerado estável quando apresenta pequenas variações no seu comportamento geral quando é avaliado em diversas condições de ambiente. Outro conceito importante é o da adaptabilidade ou capacidade de resposta à melhoria do ambiente. Esse conceito está associado à plasticidade dos genótipos e, na literatura internacional, as vezes é referido como sensibilidade ambiental dos genótipos. Nesses termos, um genótipo ideal é aquele que responde de maneira previsível ou proporcional ao estímulo ambiental.

Os procedimentos de análise da interação genótipos x ambientes evoluíram da tradicional ANOVA conjunta de experimentos, passando pelos métodos de estudo da estabilidade e adaptabilidade fenotípica baseados em análise de regressão, pelos métodos não paramétricos para estabilidade e adaptabilidade e pelos modelos multiplicativos (AMMI) para os efeitos da interação. Tais procedimentos apresentam limitação para lidar com dados desbalanceados, delineamentos experimentais não ortogonais (blocos incompletos) e com a heterogeneidade de variâncias entre os vários locais de experimentação, situações estas corriqueiras na experimentação de campo. Além do mais, tais metodologias assumem, em geral, que os efeitos de tratamentos genéticos são fixos, o que é desvantajoso e incoerente com a prática simultânea da estimação de componentes de variância e parâmetros genéticos (tais quais a herdabilidade) realizada com base nestes experimentos.

Outro aspecto refere-se à escolha do procedimento a ser aplicado, dentre os vários disponíveis. Alguns procedimentos conduzem a resultados similares, outros possuem propriedades estatísticas superiores e alguns permitem interpretações mais simples dos resultados. Assim, a escolha a priori do método a aplicar é tarefa difícil. Também vários métodos não são prontamente comparáveis, pelo menos formalmente, pois envolvem conceitos diferentes. Neste sentido, Piepho (1999) propôs o uso da metodologia de modelos mistos via REML para a comparação entre os vários procedimentos, tais quais o de Finlay e Wilkinson, o de Eberhart e Russel, o de Shukla, o de Lin et al. e uma versão AMMI considerando os efeitos de locais como aleatórios. Segundo Piepho (1999), a maioria das medidas tradicionais de estabilidade podem ser enquadradas na metodologia de modelos mistos, assumindo os efeitos de genótipos como fixos e os efeitos de locais como aleatórios (procedimento REML/GLS). O

método REML é então aplicado na estimação de parâmetros, fato que é vantajoso devido à sua aplicabilidade para a situação de dados desbalanceados e de heterogeneidade de variâncias. Adicionalmente, propicia uma escolha formal do melhor procedimento através do uso do teste da razão de verossimilhança (REMLRT), visto que os vários modelos de análise se encaixam como sub-modelos de um modelo mais geral (o modelo de variância ambiental de Lin et al.), produzindo uma estrutura hierárquica de modelos, permitindo o uso do teste da razão de verossimilhança (REMLRT). Assim, o problema de escolha de uma medida adequada de estabilidade e adaptabilidade equivale exatamente ao problema de identificação da mais apropriada estrutura de variância e covariância. Em outras palavras, a escolha do método mais adequado é dependente do conjunto de dados analisados.

Embora adequada para lidar com desbalanceamento e heterogeneidade de variâncias, a metodologia de modelos mistos é mais adequada aos propósitos do melhoramento quando considera os efeitos de genótipos como aleatórios, visando à obtenção dos BLUPs dos referidos efeitos. E isto não é realizado pelos métodos de estabilidade e adaptabilidade mencionados. Considerando os efeitos genotípicos como aleatórios, o procedimento ideal é o BLUP multivariado (Resende et al., 1999; Resende, 2002a, p.257) em que os vários ambientes são considerados como se fossem diferentes caracteres. Neste caso, são preditos valores genéticos para cada ambiente, para o ambiente médio e para novos ambientes. O BLUP multivariado considera intrinsecamente a heterogeneidade de variâncias, sendo, portanto, o procedimento ideal. Entretanto, com grande número de ambientes, o modelo multivariado é praticamente impossível de ser ajustado. Uma opção de modelo parcimonioso para o BLUP multivariado é o modelo fator analítico multiplicativo misto (FAMM), o qual é análogo ao AMMI, pois é multiplicativo, mas difere por considerar os efeitos genotípicos como aleatórios (Resende e Thompson, 2004). Os modelos FAMM permitem inferências sobre valores genéticos, adaptabilidade, estabilidade e agrupamento de locais com base na interação genótipos x ambientes. Permite também o uso de modelos espaciais para os erros dentro de locais, que são, em geral, correlacionados.

Outra opção de modelo parcimonioso é usar o modelo misto univariado de efeitos principais (g) e interação (ge), porém, levando em conta a heterogeneidade de variâncias, via alguma transformação prévia nos dados. Simulações realizadas pelo autor indicaram que a transformação dos dados, multiplicando-os por h_i/h_{im} praticamente reproduz, via g + ge, os resultados do modelo BLUP multivariado, conduzindo a um viés para o efeito genotípico médio (g + gem, em que gem é o efeito médio das interações com locais) de apenas 2 %. No caso, h_i e h_{im} referem-se às raízes quadradas das

herdabilidades no ambiente i e da média das herdabilidades em cada ambiente, respectivamente. Esta transformação considera tanto a heterogeneidade de variância genética quanto ambiental e mostrou-se superior (em termos de viés) a outras transformações comumente relatadas em literatura, as quais são baseadas apenas no desvio padrão fenotípico (heterogeneidade de variância fenotípica). Maiores detalhes sobre isso são apresentados no tópico 6.

É importante relatar que o BLUP dos efeitos ge eliminam os chamados ruídos da interação genótipos x ambientes. Isto pode ser visto considerando a predição BLUP obtida a partir de uma tabela de dupla entrada com genótipos (g) e ambiente (e), contendo as médias de cada genótipo em cada ambiente. O modelo associado a esta tabela é $Y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) + \left(\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right) + \varepsilon_{ij}$, em que ε_{ij} é o resíduo associado às médias em cada ambiente. A predição BLUP da média genotípica em cada local ($\mu + g_i + e_j + ge_{ij}$) é dada por $I = \bar{Y}_{.j} + h_g^2(\bar{Y}_{i.} - \bar{Y}_{..}) + h_{ge}^2\left(\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right)$, quando se considera os efeitos de ambiente como fixos (modelo misto) e por $I = \bar{Y}_{..} + h_g^2(\bar{Y}_{i.} - \bar{Y}_{..}) + h_e^2(\bar{Y}_{.j} - \bar{Y}_{..}) + h_{ge}^2\left(\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right)$, quando se considera os efeitos de ambiente como aleatórios (modelo aleatório). Estes índices (I) são similares aos índices multi-efeitos associados ao delineamento de blocos ao acaso com uma planta por parcela apresentados por Resende e Higa (1994). No presente caso, os ponderadores do índice são:

$h_g^2 = \dfrac{\sigma_g^2 + \sigma_{ge}^2 / L}{\sigma_g^2 + \sigma_{ge}^2 / L + \sigma_\varepsilon^2 / L}$: herdabilidade dos efeitos de genótipos.

$h_{ge}^2 = \dfrac{\sigma_{ge}^2}{\sigma_{ge}^2 + \sigma_\varepsilon^2}$: herdabilidade dos efeitos da interação g x e.

$h_e^2 = \dfrac{\sigma_e^2 + \sigma_{ge}^2 / L}{\sigma_e^2 + \sigma_{ge}^2 / G + \sigma_\varepsilon^2 / G}$: coeficiente de determinação dos efeitos de ambiente.

L e G referem-se aos números de locais e de genótipos, respectivamente, σ_ε^2 é a variância residual associada as médias Y_{ij} e σ_ε^2 é a variância entre locais.

Verifica-se por estes índices que o BLUP de ge considera a herdabilidade dos efeitos da interação g x e, ou seja, elimina os ruídos ou efeitos residuais, por ocasião do processo de predição de ge.

Predito dessa forma, $\hat{g}_i$ fornece o efeito genotípico em um ambiente médio representado pelos L locais. O efeito genotípico predito em um ambiente não avaliado (e com mesmo padrão de interação) é dado por $\hat{g}_i^* = [(\sigma_g^2/(\sigma_g^2 + \sigma_{ge}^2/L)]\hat{g}_i$, ou seja, $\hat{g}_i^* < \hat{g}_i$, a menos que L seja grande e/ou σ_{ge}^2 seja desprezível. Para o caso balanceado, os ordenamentos por $\hat{g}_i^*$ e $\hat{g}_i$ são idênticos. Os modelos BLUP que ajustam g e ge simultaneamente fornecem $\hat{g}_i^*$ e $\hat{ge}_{ij}^*$. Nesse caso, $\hat{g}_{ij}$ pode ser obtido por $\hat{g}_{ij} = \hat{g}_i^* + \hat{ge}_{ij}^*$. O BLUP multivariado fornece $\hat{g}_{ij}$. A média desses, através dos locais, fornece $\hat{g}_i$ em qualquer dos dois procedimentos mencionados. O BLUP univariado sem ajustar ge fornece $\hat{g}_i$.

Atualmente, procedimentos de interpretação mais simples têm tido apelo para a análise de estabilidade e adaptabilidade. Neste sentido, medidas que incorporam ambos (estabilidade e adaptabilidade) em uma única estatística, tais quais os métodos de Annicchiarico (1992) e Lin e Binns (1988) e modificações, têm sido enfatizados (Cruz e Carneiro, 2003). No contexto dos modelos mistos, um método para ordenamento de genótipos simultaneamente por seus valores genéticos (produtividade) e estabilidade, refere-se ao procedimento BLUP sob médias harmônicas (Resende, 2002a, p. 344). Neste caso, o vetor de dados (y) deve ser trabalhado como a recíproca dos dados observados, ou seja, (1/y). Isto produz resultados que são função (1/H) da média harmônica (H) dos dados. Quanto menor for o desvio padrão do comportamento genotípico através dos locais, maior será a média harmônica de seus valores genotípicos através dos locais. Assim, a seleção pelos maiores valores da média harmônica dos valores genotípicos (MHVG) implica simultaneamente seleção para produtividade e estabilidade. Adicionalmente, o uso da transformação 1/y considera também a instabilidade dentro dos locais.

Em termos de adaptabilidade, uma medida simples e eficiente no contexto dos modelos mistos refere-se à performance relativa dos valores genotípicos (PRVG) através dos ambientes. Neste caso, os valores genotípicos preditos (ou os dados originais) são expressos como proporção da média geral de cada local (ML) e, posteriormente, obtém-se o valor médio desta proporção através dos locais. Genericamente, a performance relativa tem sido usada há longo tempo (Wright et al. 1966) em termos de dados fenotípicos, e constitui a base do método de Annicchiarico (1992).

A seleção simultaneamente por produtividade, estabilidade e adaptabilidade, no contexto dos modelos mistos, pode ser realizada pelo método da média harmônica da performance relativa dos

valores genéticos (MHPRVG) preditos, proposto por Resende (2004). Este método permite selecionar simultaneamente pelos três atributos mencionados e apresenta as seguintes vantagens: (i) considera os efeitos genotípicos como aleatórios e portanto fornece estabilidade e adaptabilidade genotípica e não fenotípica; (ii) permite lidar com desbalanceamento; (iii) permite lidar com delineamentos não ortogonais; (iv) permite lidar com heterogeneidade de variâncias; (v) permite considerar erros correlacionados dentro de locais; (vi) fornece valores genéticos já descontados (penalizados) da instabilidade; (vii) pode ser aplicado com qualquer número de ambientes; (viii) permite considerar a estabilidade e adaptabilidade na seleção de indivíduos dentro de progênie; (ix) não depende da estimação e interpretação de outros parâmetros tais quais coeficientes de regressão; (x) elimina os ruídos da interação genótipos x ambientes pois considera a herdabilidade desses efeitos; (xi) gera resultados na própria grandeza ou escala do caráter avaliado; (xii) permite computar o ganho genético com a seleção pelos três atributos simultaneamente. Estes dois últimos fatores são bastante importantes. Outros métodos como o de Lin e Binns fornecem resultados que não são interpretados diretamente como valores genéticos e então não permitem computar o ganho genético no caráter composto pela produtividade, estabilidade e adaptabilidade. O método de Annicchiarico depende, adicionalmente, de suposições de valores de α associados a $Z_{(1-\alpha)}$, que refere-se ao percentil da função distribuição normal padrão associado a determinado nível de significância α.

Em resumo, considerando os efeitos de genótipos como aleatórios, existem duas opções principais de análise via modelos mistos REML/BLUP: (i) modelos FAMM, os quais são análogos aos modelos AMMI; (ii) MHPRVG, que é análogo aos métodos de Lin e Bins e Annicchiarico. A MHPRVG pode ser aplicada via modelo multivariado ou via modelo univariado do tipo g + ge com correção para heterogeneidade de variâncias.

No presente texto, somente essas duas metodologias serão descritas. Vários métodos para estudo da adaptabilidade e estabilidade fenotípica no contexto dos modelos com efeitos fixos de tratamento são descritos com detalhes nas obras de Vencovsky e Barriga (1992), Ramalho et al. (1993), Cruz e Regazzi (1994), Duarte e Vencovsky (1999) e Cruz e Carneiro (2003).

3 MÉTODO MHPRVG SIMULTÂNEO PARA PRODUTIVIDADE, ESTABILIDADE E ADAPTABILIDADE

Na Tabela 27 são apresentados alguns resultados comparativos dos métodos MHPRVG, Lin e Bins e Annicchiarico, envolvendo cinco clones (C) avaliados em 6 locais (L).

Tabela 27. Resultados comparativos envolvendo as medidas de produtividade via valores genotípicos (VGMed), produtividade e estabilidade via valores genotípicos (MHVG), produtividade e adaptabilidade via valores genotípicos (PRVG), produtividade, estabilidade e adaptabilidade via valores genotípicos (MHPRVG), produtividade, estabilidade e adaptabilidade via métodos de lin e Bins (Pi) e Annicchiarico (Annicch.). Avaliação de cinco clones (C) em seis locais (L), caráter circunferência do tronco em uma espécie florestal.

	L1	L2	L3	L4	L5	L6	VGMed	MHVG
				Valores Genotípicos Preditos para Circunferência				
C1	28.3008	28.5118	30.1782	30.7418	28.8242	30.725	29.5469	29.511
C2	30.5457	28.4889	28.0309	27.0569	30.1967	**32.946**	29.5442	29.420
C3	30.7785	27.4835	31.0735	32.1535	30.1863	25.810	29.5810	29.407
C4	31.0119	**34.1820**	**33.6280**	**33.3820**	30.3450	28.755	31.8840	31.759
C5	**32.2383**	31.9302	32.1674	30.6400	**32.1403**	31.608	31.7874	31.778
Me	30.5750	30.1193	31.0156	30.7948	30.3385	29.969	30.4687	30.375

	L1	L2	L3	L4	L5	L6	PRVG	MHPRVG
				Performance Relativa dos Valores Genotípicos Preditos em Relação à Média do Local				
C1	0.92562	0.94663	0.97300	0.99828	0.95009	1.02523	0.96981	0.96866
C2	0.99904	0.94587	0.90377	0.87862	0.99533	**1.09934**	0.97033	0.96507
C3	1.00666	0.91249	1.00187	1.04412	0.99498	0.86122	0.97022	0.96597
C4	1.01429	**1.13489**	**1.08423**	**1.08401**	1.00021	0.95950	1.04619	1.04278
C5	**1.05440**	1.06012	1.03714	0.99497	**1.05939**	1.05471	1.04346	1.04293

Ord	VGMed	MHVG	PRVG	MHPRVG	Pi	Annicch	Pi	Annicch.
1	C5	C5	C4	C5	C5	C5	6.871	0.95975
2	C4	C4	C5	C4	C4	C4	9.199	0.94862
3	C1	C1	C2	C1	C1	C1	9.149	0.95143
4	C2	C2	C3	C3	C3	C3	1.858	1.02831
5	C3	C3	C1	C2	C2	C2	1.376	1.03657

Me = média; Ord = ordenamento. Valores em negrito referem-se aos maiores valores observados em cada local.

Verifica-se que a estatística MHPRVG produziu exatamente o mesmo ordenamento que as estatísticas de Lin e Binns (Pi) e de Annicchiarico e pode ser usada vantajosamente no contexto dos modelos mistos com efeitos genéticos aleatórios. Ressalta-se que aqui as metodologias de Lin e Binns (Pi) e de Annicchiarico foram aplicadas sobre valores genotípicos preditos por BLUP. A aplicação das mesmas sobre médias fenotípicas, conforme as metodologias originais pode gerar resultados menos precisos.

A MHPRVG deve ser aplicada preferencialmente sobre os dados originais, expressando-os como (média do local)/y e posteriormente obtendo-se os BLUPs para os valores genotípicos (média geral + efeitos genotípicos). A recíproca destes, multiplicada pela média geral de todos os ensaios, fornece a MHPRVG na unidade de avaliação do caráter. Procedendo-se desta forma, as diferentes precisões associadas aos valores genéticos preditos dos genótipos nos ambientes são automaticamente levadas em consideração pelo procedimento REML/BLUP.

A seguir, descrevem-se resultados da avaliação de 180 clones de cana-de-açúcar em três locais no Estado do Paraná, no contexto do Programa de Melhoramento da Cana-de-Açúcar da UFPR (Oliveira et al., 2005). Para o delineamento de blocos ao acaso com uma observação por parcela e em vários ambientes ou locais, adotou-se o seguinte modelo estatístico:

y = Xb + Zg + Wgl + e, em que:

y, b, g, gl, e = vetores de dados, de efeitos fixos (médias de blocos através dos locais), de efeitos genotípicos de clones (aleatório), de efeitos da interação genótipos x ambientes (aleatório) e de erros aleatórios, respectivamente.

X, Z e W = matrizes de incidência para b, g e gl, respectivamente.

Distribuições e estruturas de médias e variâncias:

$$E\begin{bmatrix} y \\ g \\ gl \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \\ 0 \end{bmatrix}; \qquad Var\begin{bmatrix} g \\ gl \\ e \end{bmatrix} = \begin{bmatrix} I\sigma_g^2 & 0 & 0 \\ 0 & I\sigma_{gl}^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix}$$

Esse modelo ajusta os efeitos de locais e de blocos dentro de locais no vetor dos efeitos

fixos por meio da combinação bloco-local, a qual contempla todos os graus de liberdade disponíveis nas fontes de variação referentes a locais e blocos dentro de locais.

Os valores genotípicos preditos para o clone i em cada local j usa simultaneamente os dados de todos os locais e são dados por $VG_{ij} = u_j + \hat{g}_i + \hat{gl}_{ij}$, em que u_j é a média do local j. Nesse caso, tanto g quanto gl são preditos com maior precisão pois todo o conjunto de dados é usado bem como os ruídos da interação são eliminados quando se produzem os BLUP´s de gl. Na Tabela 28 estão apresentados os resultados sobre a estabilidade (MHVG – Média Harmônica dos Valores Genotípicos através dos locais), adaptabilidade (PRVG – Performance Relativa dos Valores Genotípicos em relação a média de cada local), e estabilidade e adaptabilidade simultaneamente (MHPRVG – Média Harmônica da Performance Relativa dos Valores Genotípicos), para o caráter TCH (t ha^{-1}), para os 20 melhores clones dentre aqueles avaliados em todos os locais.

Verifica-se (Tabela 28) que os dez melhores clones, com base nos critérios PRVG, MHVG e MHPRVG não são exatamente os dez melhores pelo critério de produtividade média via valores genotípicos (Tabela 29). A coincidência foi de 80 % dentre os dez melhores e houve inversão de ordem dentre os coincidentes. Isto evidencia que a utilização desses novos atributos ou critérios de seleção podem propiciar um refinamento a mais na seleção. Os dois melhores clones pelo critério MHPRVG apresentaram superioridade média de 28 % (RB955466) e 19 % (RB965518) sobre a média geral dos três locais. Estes valores foram computados já penalizando os clones pela instabilidade através dos locais e ao mesmo tempo capitalizando a capacidade de resposta (adaptabilidade) à melhoria do ambiente. Essas propriedades são intrínsecas ao método MHPRVG. Os valores de PRVG e MHPRVG na Tabela 28 indicam exatamente a superioridade média do genótipo em relação à média do ambiente em que for cultivado. Assim, o genótipo RB955466 responde em média 1,28 vezes a média do ambiente em que for plantado. O valor de MHPRVG*MG fornece o valor genotípico médio dos clones nos locais avaliados, valor este já penalizado pela instabilidade e capitalizado pela adaptabilidade.

Os cinco melhores clones a serem selecionados com base no método da MHPRVG são: RB955466, RB965518, RB965648, RB965718 e RB965743. Tal seleção propicia um ganho de 19,80 % sobre a média geral dos três ambientes, considerando simultaneamente a produtividade, estabilidade e adaptabilidade através dos locais. Na Tabela 29 são apresentados (somente os 20

melhores clones dentre aqueles avaliados em todos os locais) resultados referentes à seleção simultânea por produtividade, adaptabilidade e estabilidade por meio do emprego do método de Lin e Binns (1988) sobre os valores genotípicos preditos. Na mesma Tabela são apresentados os valores genotípicos capitalizando a interação média (gem) nos vários locais.

Verifica-se que dentre os dez melhores clones selecionados pela MHPRVG, nove coincidem com os dez clones selecionados pelo método de Lin e Binns (1988). Nesse método, os melhores materiais genéticos são aqueles com menores valores da estatística Pi. Tomando por base o método de Lin e Binns (1988), o genótipo não coincidente foi o RB966200, o qual ficou em décimo lugar no ordenamento por esse método e em décimo terceiro no método MHPRVG, portanto, em posições muito próximas. A correlação estimada entre os parâmetros dos dois métodos foi de alta magnitude (–0,9487). O método de Annicchiarico (1992) foi também computado e apresentou correlação absoluta nessa mesma magnitude (porém positiva) com o método MHPRVG. Isso confirma que os três métodos utilizam basicamente os mesmos princípios e conceitos. O método MHPRVG apresenta a vantagem de fornecer resultados na própria escala de medição do caráter, os quais podem ser interpretados diretamente como valores genéticos para o caráter avaliado. Isto permite também calcular o ganho genético com a seleção simultânea para produtividade, adaptabilidade e estabilidade. Isto não é possível com o método de Lin e Binns (1988). Assim, a estatística MHPRVG pode ser usada vantajosamente no contexto dos modelos mistos com efeitos genéticos aleatórios. A consideração dos efeitos genéticos e da interação g x e como aleatórios propicia vantagem também sobre o método AMMI (Gauch, 1988), o qual trata esses efeitos como fixos e, portanto, atua no nível fenotípico e não genotípico. O BLUP dos efeitos da interação já elimina os ruídos de tais efeitos, à semelhança do método AMMI, conforme relatado por Resende (2004). É importante relatar, no entanto, que o aspecto multiplicativo da interação na filosofia AMMI é muito importante e mais completa. O uso desse aspecto, porém, sob efeitos aleatórios de genótipos, produz a técnica FAMM, a qual é vantajosa sobre a AMMI e é detalhada em tópico seguinte.

Comparação entre as várias Predições de Valores Genotípicos

Seis modalidades de valores genotípicos foram preditas para cada clone: (a) por local (u + g + gl, Tabela 30); (b) para vários locais, livres da interação g x e (u + g, Tabela 30); (c) para a média

dos locais, capitalizando o efeito médio da interação (u + g + gem, Tabela 29); (d) para vários locais penalizando pela instabilidade de cada genótipo (MHVG, Tabela 28); (e) para a média dos locais, capitalizando a capacidade de resposta de cada genótipo à melhoria do ambiente (PRVG, Tabela 28); (f) para a média dos locais, penalizando pela instabilidade e capitalizando pela adaptabilidade (MHPRVG, Tabela 28).

Em termos de inferências sobre a produtividade esperada, tais valores genotípicos devem ser usados da seguinte maneira:

(i) para plantio em cada local da rede experimental: considerar valores genotípicos (médias genéticas) descritos em (a);

(ii) para plantio em vários outros locais com o mesmo padrão de interação g x e da rede experimental: considerar valores genotípicos (médias genéticas) descritos em (c) ou (e);

(iii) para plantio em outros locais desconhecidos ou com padrão de interação g x e diferente daquele da rede experimental ou com alta heterogeneidade ambiental dentro de local: considerar valores genotípicos (médias genéticas) descritos em (b) ou (d);

(iv) para plantio em vários outros locais com variados padrão de interação g x e: considerar valores genotípicos (médias genéticas) descritos em (f).

Os métodos que mais penalizam os valores genotípicos preditos são, pela ordem: (d) e (b); (f), (e) e (c); (a). Dentre esses, (d) e (b) são similares, sendo que (d) tende a ser superior, na consideração do conceito de estabilidade, por particularizar melhor a interação para cada genótipo. Também (c), (e) e (f) geram resultados mais similares entre eles. De maneira genérica, pode-se dizer que os métodos MHVG e MHPRVG são opções seguras, sendo o MHVG um pouco mais conservador. No presente exemplo, dentre os genótipos avaliados em todos os locais, o clone RB955466 foi o primeiro colocado em todos os critérios: produtividade, estabilidade e adaptabilidade e os três atributos simultaneamente. Nas demais posições, houve certa alternância de genótipos de acordo com o critério (Tabelas 28 e 29).

Tabela 30. Valores genotípicos de 22 genótipos de cana-de-açúcar em estudo e ganhos genéticos preditos dos cinco melhores para o caráter TCH (toneladas de cana por hectare), em três ambientes, e na análise conjunta, no Estado do Paraná. O somatório $u_j + g + gl_j$ equivale à média do local j (u_j) somada aos efeitos de genótipos (g) e da interação genótipos x local j (ge_j).

Clones	Ambiente 1 (Colorado)		Clones	Ambiente 2 (Paranavaí)		Clones	Ambiente 3 (Mandaguaçú)		*Clone*	Análise Conjunta	
	Valores genéticos	Ganho genético		Valores genéticos	Ganho genético		Valores genéticos	Ganho genético		Valores genéticos	Ganho genético
	$u_1 + g + gl_1$	(%)		$u_2 + g + gl_2$	(%)		$u_3 + g + gl_3$	(%)		(u + g)	(%)
RB955466	96,88	32,68	RB955466	132,31	41,43	RB965602	192,46	26,06	RB945273	130,78	22,86
RB965560	96,64	32,52	RB965674	115,77	32,57	RB945273	192,07	25,93	RB955466	130,00	22,49
RB965602	96,37	32,34	RB965648	115,71	29,60	RB965689	177,83	22,78	RB965602	129,60	22,24
RB965699	91,83	30,69	RB965718	114,67	27,83	RB965564	173,49	20,50	RB965731	129,32	22,05
RB965674	91,06	29,49	RB965518	112,06	26,22	RB965518	173,48	19,12	RB965718	125,76	21,27
RB965518	89,72		RB965560	110,95		RB966256	173,15		RB965648	123,40	
RB965648	89,06		RB965741	110,93		RB955466	172,97		RB965518	122,93	
RB965688	88,98		RB965689	110,79		RB882698	172,70		RB965743	122,70	
RB965743	88,67		RB965657	110,39		RB965743	171,24		RB965674	121,51	
RB965718	88,44		RB965743	110,23		RB965731	170,32		RB965699	121,40	
RB882698	88,07		RB966200	107,78		RB965698	169,46		RB965689	121,18	
RB965657	87,80		RB893161	106,18		RB965648	169,24		RB965560	120,34	
RB965741	86,01		RB892677	105,91		RB965624	169,12		RB882698	119,36	
RB966200	85,01		RB965699	105,80		RB965591	168,78		RB965591	118,74	
RB965564	84,24		RB965688	105,71		RB965574	168,67		RB966256	118,54	
RB965689	84,06		RB965574	105,69		RB966200	168,61		RB965657	118,47	
RB965641	84,05		RB882698	105,31		RB965657	167,92		RB965625	118,14	
RB965675	83,99		RB965658	104,59		RB965586	167,27		RB965741	117,95	
RB892677	83,95		RB965614	104,50		RB965517	166,93		RB966200	117,83	
RB966256	83,94		RB965546	104,33		RB965741	166,59		RB965688	117,54	
RB72454*	78,21		RB72454*	100,45		RB72454*	161,63		RB72454*	112,23	
RB835486*	66,57		RB835486*	86,57		RB835486*	146,57		RB835486*	102,15	

* Padrões: cultivares em uso no Estado do Paraná

Tabela 28. Estabilidade de valores genotípicos (MHVG), adaptabilidade de valores genotípicos (PRVG), estabilidade e adaptabilidade de valores genotípicos (MHPRVG) para TCH (toneladas de cana por hectare) dos 20 melhores clones e mais duas testemunhas.

Genótipo	MHVG	Genótipo	PRVG	PRVG*MG	Genótipo	MHPRVG	MHPRVG*MG
RB955466	**126,79**	**RB955466**	**1,29**	**137,39**	**RB955466**	**1,28**	**136,21**
RB965560	**116,71**	**RB965648**	**1,19**	**126,43**	**RB965518**	**1,19**	**126,23**
RB965648	**116,37**	**RB965518**	**1,19**	**126,36**	**RB965648**	**1,18**	**126,14**
RB965518	**116,13**	**RB965560**	**1,18**	**125,62**	**RB965718**	**1,17**	**124,58**
RB965718	**115,13**	**RB965718**	**1,17**	**124,95**	**RB965743**	**1,17**	**124,50**
RB965743	**114,55**	**RB965743**	**1,17**	**124,63**	**RB965560**	**1,17**	**124,34**
RB965657	**113,63**	**RB965689**	**1,17**	**124,14**	**RB965689**	**1,17**	**124,12**
RB965689	**113,01**	**RB965657**	**1,16**	**123,50**	**RB965657**	**1,16**	**123,33**
RB965699	**112,87**	**RB882698**	**1,15**	**122,82**	**RB882698**	**1,15**	**122,70**
RB882698	**112,61**	**RB965741**	**1,15**	**122,53**	**RB965741**	**1,15**	**122,36**
RB965741	112,60	RB965699	1,15	121,95	RB965699	1,14	121,29
RB965688	112,08	RB965688	1,14	121,58	RB965688	1,14	121,28
RB966200	111,23	RB966200	1,14	121,32	**RB966200**	1,14	121,26
RB966256	109,31	RB966256	1,13	119,86	RB966256	1,13	119,81
RB965658	108,92	RB965574	1,12	119,31	RB965574	1,12	119,29
RB965591	108,39	RB965564	1,12	118,91	RB965564	1,12	118,76
RB965564	108,33	RB965658	1,11	118,70	RB965658	1,11	118,63
RB966215	107,61	RB965591	1,11	118,59	RB965591	1,11	118,57
RB893161	107,14	RB965625	1,10	116,99	RB965625	1,10	116,91
RB965625	107,13	RB965614	1,10	116,92	RB966215	1,10	116,70
RB72454*	103,70	RB72454*	1,07	113,62	RB72454*	1,07	113,61
RB835486*	91,61	RB835486*	0,95	100,89	RB835486*	0,95	100,87

* Padrões: cultivares em uso no Estado do Paraná

Tabela 29. Valores genotípicos, capitalizando a interação média (gem) nos vários locais, bem como estabilidade e adaptabilidade de valores genotípicos para TCH (toneladas de cana por hectare) por meio do método (Pi) de Lin e Binns (1988), em que $Pi = \sum_j (VG_{ij} - M_j)^2/(2L)$, VG_{ij} é o valor genotípico do genótipo i no local j, M_j é o valor genotípico máximo no local j e L é o número de locais.

Genótipo	u + g + gem	Genótipo	Pi
RB955466	**134,87**	**RB955466**	**63,34**
RB965718	**129,76**	**RB965518**	**136,92**
RB965648	**126,91**	**RB965689**	**140,29**
RB965518	**126,34**	**RB965648**	**146,00**
RB965743	**126,07**	**RB965743**	**167,58**
RB965699	**124,49**	**RB965718**	**181,94**
RB965689	**124,22**	**RB965657**	**194,15**
RB965560	**123,22**	**RB882698**	**199,46**
RB882698	**122,03**	**RB965741**	**207,35**
RB965591	**121,29**	**RB966200**	**218,54**
RB966256	121,04	RB966256	238,00
RB965657	120,96	RB965574	247,05
RB965625	120,56	RB965688	255,73
RB965741	120,34	RB965564	266,88
RB966200	120,19	RB965591	268,14
RB965688	119,83	RB965560	277,93
RB965574	119,66	RB965658	282,70
RB965564	119,53	RB965699	293,65
RB955452	118,74	RB965698	320,17
RB965658	118,07	RB965625	320,81
RB72454*	113,43	RB72454*	385,63
RB835486*	101,26	RB835486*	798,76

* Padrões: cultivares em uso no Estado do Paraná.

De maneira geral os experimentos instalados em vários locais são avaliados também em várias colheitas (culturas perenes) e em vários anos de plantio (culturas anuais). A análise desse tipo de rede experimental é discutida em detalhes no tópico 10.3 do Capítulo 10 (culturas perenes) e Resende (2007) (culturas anuais).

Nesse último caso, o seguinte modelo linear misto pode ser ajustado.

y = Xf + Zg + Qga + Tgl + Wgla + e, em que:

y é o vetor de dados, f é o vetor dos efeitos das combinações repetição-local-ano (assumidos como fixos ou aleatórios) somados à média geral, g é o vetor dos efeitos genotípicos (assumidos como aleatórios), ga é vetor dos efeitos da interação de genótipos com anos (aleatórios), gl é o vetor dos efeitos da interação genótipos x locais, gla é o vetor dos efeitos da interação tripla genótipos x locais x anos (assumidos como aleatórios) e e é o vetor de erros ou resíduos (aleatórios). As letras maiúsculas representam as matrizes de incidência para os referidos efeitos.

Esse modelo é equivalente ao utilizado tradicionalmente para a análise envolvendo a avaliação de genótipos em vários locais e anos de plantio, dado por $Y_{ijkn} = \mu + g_i + b_{j/k/n} + a_k + l_n + ga_{ik} + gl_{in} + al_{kn} + gal_{ikn} + gbal_{ij/k/n}$, em que u é o efeito da média geral, gi é o efeito do genótipo i, $b_{j/k/n}$ é o efeito do bloco j dentro do ano k dentro do local n, l_n é o efeito do local n, a_k é o efeito do ano de plantio k, gl_{in} é o efeito da interação genótipos x locais, ga_{ik} é o efeito da interação genótipos x anos de plantio, al_{kn} é o efeito da interação locais x anos de plantio, gal_{ikn} é o efeito da interação genótipos x anos x locais, e $gbal_{ij/k/n}$ é o erro ou resíduo aleatório. No caso, os efeitos de anos, locais e blocos/anos/locais foram agrupados no efeito f, pois são efeitos puramente ambientais. Nesse caso, é essencial que as repetições sejam codificadas com diferentes números nos diferentes locais e anos. Modelos desse tipo foram usados por Smith et al. (2001), na Austrália, e podem ser ajustados pelos modelos 114 e 125 do Selegen-Reml/Blup. Modelos como os apresentados no presente tópico, quando ajustados via metodologia de modelos mistos (procedimento REML/BLUP), são muito vantajosos, sobretudo pela facilidade e eficiência estatística com que permitem lidar com dados incompletos, por exemplo, para o caso em que nem todas as variedades encontram-se em todos os experimentos.

4 MÉTODO FAMM PARA ANÁLISE ESTATÍSTICA DA INTERAÇÃO GENÓTIPOS X AMBIENTES VIA MODELOS MISTOS

Os métodos tradicionais de análise da interação genótipos x ambientes baseados em ANOVA e regressão linear apresentam limitações tais quais a falta de habilidade para explicar grande parte dessa interação, a não informação sobre interações específicas positivas ou negativas com ambientes, a falta de linearidade nas relações genótipo – ambiente em alguns casos e a falta de decomposição da interação em termos padrão ou tendências e ruídos (Duarte e Vencovsky, 1999). Uma primeira tentativa para contornar essas limitações foi a proposição da técnica AMMI (efeitos aditivos principais e multiplicativos da interação genótipos x ambientes). Essa técnica foi indicada para uso na experimentação agronômica por Kempton (1984) na Inglaterra e por Brian (1978) na França (ver Gallais, 1989, página 77). Posteriormente foi bem descrita por Gauch (1988; 1992). É atribuída a Fisher e Mackenzie (1923) e Gollob (1968). Outra denominação do método é Análise de Componentes Principais (PCA) Duplamente Centrada, já que envolve a decomposição da matriz de interações em termos de autovalores e autovetores. Kempton (1984) relata a resposta de um cultivar como uma série de termos multiplicativos, cada termo sendo um produto efeito genótipo x efeito do ambiente. O procedimento AMMI pode ser visto como um método para separar o padrão (isto é, o efeito da interação g x e propriamente dito) do ruído (erro médio da média de tratamento dentro de ensaio). Isto é conseguido por PCA, onde os primeiros eixos (isto é, os eixos com os maiores autovalores), recuperam a maioria do padrão, enquanto a maioria do ruído é captada pelos últimos eixos. O padrão pode ser visto como todo o efeito da interação g x e ponderado por uma estimativa da proporção padrão/ruído associada com o respectivo efeito. Essa proporção é uma razão entre componentes de variância, análoga a um coeficiente de herdabilidade ou repetibilidade (Piepho, 1994).

Os modelos AMMI foram popularizados no contexto de modelos de efeitos fixos e foram aplicados em várias oportunidades (Gauch, 1988; 1992; Crossa et al., 1990; Duarte e Vencovsky, 1999). A análise AMMI combina em um modelo, componentes aditivos para os efeitos principais (genótipos e ambientes) e componentes multiplicativos para os efeitos g x e. Então combina a técnica univariada da ANOVA para os efeitos principais e uma técnica multivariada (PCA) para os

efeitos g x e. Crossa (1990) sugere que o uso de técnicas multivariadas permite um melhor uso da informação do que os tradicionais métodos de regressão.

Embora úteis, os modelos AMMI apresentam pelo menos cinco grandes limitações: consideram os efeitos de genótipos e de g x e como fixos; é adequado somente para dados balanceados; não permite considerar a variação espacial dentro de ensaios; não considera a heterogeneidade de variância entre ensaios; não considera os diferentes números de repetições entre ensaios. Esses fatores não são realísticos na análise de experimentos de campo, em que os dados são geralmente desbalanceados e os tratamentos (genótipos) não suportam a suposição de efeitos fixos (implicitamente herdabilidade igual a 1) e devem ser tratados como efeitos aleatórios (ver Capítulo 3 e a discussão a seguir). Assim, os modelos AMMI estimam valores fenotípicos e não valores genéticos. Se os efeitos de genótipos são considerados aleatórios, os mesmos podem ser preditos por BLUP. Hill e Rosenberger (1985), Stroup e Mulitze (1991), Resende (1999; 2002) e Smith et al. (2001a) mostraram que, assumindo genótipos como efeitos aleatórios, é preferível em termos de acurácia preditiva, mesmo quando tais genótipos são considerados como de efeitos fixos pelos padrões tradicionais.

Smith et al. (2005) relatam que se o objetivo é ter estimativas dos efeitos genotípicos tão próximas quanto possível dos verdadeiros efeitos, genótipos devem ser considerados como aleatórios. E isto independe da forma de amostragem dos genótipos em teste. Muitos geneticistas e estatísticos consideram que variedades em fase final de teste não representam uma amostra aleatória e, portanto, devem ser consideradas como de efeito fixo. Smith et al. (2005) e o presente autor não aderem a essa filosofia. A adoção dos efeitos de cultivares como aleatórios não é ainda de uso generalizado e as razões para isso são históricas. Assim, uma grande mudança de cultura é necessária (Smith et al., 2005).

Assumindo genótipos como efeitos aleatórios, é possível obter predições regressadas dos efeitos aleatórios de g x e, e então separar padrão e ruído como fazem os modelos AMMI. Nesse sentido, BLUP e AMMI podem ser vistos como duas abordagens para atingir o mesmo objetivo de separar padrão e ruído. O procedimento BLUP obtém estimativas GLS (quadrados mínimos generalizados) dos efeitos da interação e então pondera-os por uma estimativa da correspondente proporção dos componentes de variância padrão/ruídos. Entretanto, o procedimento BLUP tem

várias vantagens que contornam todas as limitações do AMMI. Piepho (1994) mostrou que o BLUP é preditivamente mais acurado do que o AMMI.

Os modelos BLUP descritos no tópico 3 desse capítulo tem grande habilidade para explicar a interação, para informar sobre interações específicas positivas ou negativas com ambientes e para decompor a interação em termos padrão ou tendências e ruídos. No entanto, assumem uma estrutura de variância de simetria composta (ver tópico 5) ou simetria composta com variâncias heterogêneas. Essa suposição pode não ser adequada quando existe grande heterogeneidade de covariância genética entre pares de locais. O modelo BLUP multivariado é a melhor opção para análise de múltiplos experimentos pois considera tanto a heterogeneidade de variância quanto de covariância, além de propiciar inferência da resposta genotípica em cada ambiente. No entanto, com grande número de ambientes, a análise de modelos mistos multivariados é superparametrizada e de difícil convergência. A estrutura de variância e covariância nesse caso é completamente não estruturada, significando que um grande número de parâmetros deve ser estimado. Então, a estrutura parcimoniosa associada ao AMMI é uma característica interessante. Van Eeuwijk et al. (1995) sugeriram obter os BLUP's das combinações genótipos x ambientes e então submeter essa tabela de dupla entrada a uma análise AMMI, usando o procedimento da decomposição por valor singular ou a PCA. Uma abordagem melhor foi encontrada por Piepho (1998), que apresentou, no contexto dos modelos mistos, um modelo fator analítico multiplicativo com efeitos aleatórios de genótipos e de g x e, o qual é conceitualmente e funcionalmente melhor que o AMMI. No mesmo contexto, Smith, Cullis e Thompson (2001b) apresentaram uma classe geral de modelos fator analíticos multiplicativos mistos que incorporam a abordagem de Piepho (1998) e inclui erros espaciais separados para cada experimento. Tais modelos propiciam uma abordagem realista e completa para a análise de múltiplos experimentos repetidos em vários ambientes (Thompson et al., 2003). Resende e Thompson (2003; 2004) denominaram tal técnica como FAMM (modelos fator analíticos mistos) e ressaltam o poder da mesma em aproximar os resultados obtidos por um modelo multivariado completo com matriz de variância e covariância não estruturada. A técnica FAMM é análoga ao AMMI, podendo ser referida como AMMI com efeitos aleatórios de genótipos e de g x e, fato que conduz a uma estrutura fator analítica.

A técnica multivariada da análise de fatores (Lawley e Maxwell, 1971; Mardia, Kent e Bibby, 1988; Comrey e Lee, 1992) propicia simplificação de dados multivariados correlacionados assim como o faz outras técnicas multivariadas, tais quais a PCA e a transformação canônica. Estas técnicas consideram as correlações entre variáveis e geram um novo conjunto de variáveis não correlacionadas. A técnica AF pode ser considerada como uma extensão da PCA e a estrutura de variância-covariância fator analítica propicia uma aproximação da estrutura multivariada não estruturada, gerando modelos parcimoniosos. Detalhes dessas técnicas são apresentados no Capítulo 7.

No contexto dos modelos mistos, modelos multiplicativos para a interação g x e induzem correlações entre as interações. Modelos mistos com termos multiplicativos são intimamente relacionados à estrutura de covariância fator analítica proposta por Jennrich e Schluchter (1986). Piepho (1997) propôs também modelos multiplicativos mistos para análise de múltiplos experimentos, mas assumiu efeitos fixos de genótipos e aleatórios de ambiente. O mesmo autor propôs posteriormente considerar efeitos de genótipos como aleatórios (Piepho, 1998). Nos modelos FAMM descritos por Smith et al. (2001b) e Resende e Thompson (2003; 2004), o ajuste dos modelos, ou seja, o número de termos multiplicativos necessários pode ser formalmente testado por meio do teste da razão de verossimilhança (REMLLRT). Por meio de uma abordagem de modelos mistos unificada, os parâmetros de estabilidade e adaptabilidade são integrados em inferências amplas (seleção para um ambiente médio), específicas (seleção para um ambiente específico) e para um novo ambiente (seleção para um ambiente não avaliado). Adicionalmente, estimativas e predições dos carregamentos ambientais e escores genotípicos são obtidas e dispostas em gráficos usando ferramentas como os *biplots*, de maneira a se entender melhor a interação g x e.

Modelos Fator Analíticos

Um modelo associado à avaliação de vários tratamentos ou genótipos em vários ambientes é dado por:

$Y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij}$, em que:

μ, *g, e, ge* e ε são os efeitos da media geral, do genótipo i, do ambiente j, da interação do genótipo i com o ambiente j e dos erros aleatórios, respectivamente. O efeito da media geral é fixo, o efeito de ambiente pode ser considerado como fixo ou alatório e os demais efeitos são considerados como aleatórios. Um modelo referente aos efeitos aleatórios de genótipos em cada ambiente pode ser escrito como: $Y_{ij} = \mu + g_{ij} + e_j + \varepsilon_{ij}$.

No contexto da análise de múltiplos experimentos, a abordagem da análise de fatores pode ser usada para propiciar uma classe de estruturas para a matriz de variância e covariância G_0, associada aos efeitos g_{ij}. O modelo de análise é postulado em termos de efeitos genotípicos não observáveis em diferentes ambientes:

$g_{ij} = \sum_{r=1}^{k} \lambda_{jr} f_{ir} + \delta_{ij}$, em que:

g_{ij} : efeito do genótipo i no ambiente j;

λ_{jr} : carregamento do fator r no ambiente j;

f_{ir} : escore para o genótipo i no fator r;

δ_{ij} : erro representando a falta de ajuste do modelo.

O modelo FAMM é apresentado com base em Smith, Cullis e Thompson (2001b) e Resende e Thompson (2003; 2004). Aplicado a g genótipos em s ambientes, o modelo fator analítico postula dependência em um conjunto de fatores hipotéticos aleatórios $f_r^{(g\,x\,1)}, (r = 1....k < s)$. Em notação vetorial, o modelo fator analítico para esses efeitos é $g_s = (\lambda_1 \otimes I_g) f_1 + ... + (\lambda_k \otimes I_g) f_k + \delta$, em que:

$\lambda_r^{(s\,x\,1)}$: carregamentos ou pesos dos fatores nos ambientes;

$\delta^{(gs\,x\,1)}$: vetor de resíduos ou falta de ajuste do modelo (também denominado vetor de fatores específicos).

De maneira compacta, o modelo é:

$g_s = (\Lambda \otimes I_g) f + \delta$, em que:

$\Lambda^{(s\,x\,k)} = [\lambda_1 ..... \lambda_k]$;

$f^{(gk\,x1)} = (f_1^{'}, f_2^{'}.... f_k^{'})'.$

A distribuição conjunta de f e δ é dada por

$\begin{pmatrix} f \\ \delta \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_k \otimes I_g & 0 \\ 0 & \Psi \otimes I_g \end{pmatrix}\right]$, em que:

$\Psi = diag\,(\psi_1.....\psi_p)$;

ψ_i : variância específica para o i-ésimo ensaio.

A matriz de variância para os efeitos dos genótipos nos ambientes é dada por

$$\begin{aligned} \text{var}(g_s) &= (\Lambda \otimes I_g)\,\text{var}(f)\,(\Lambda^{'} \otimes I_g) + \text{var}(\delta) \\ &= (\Lambda\Lambda^{'} + \Psi) \otimes I_g. \end{aligned}$$

O modelo para os efeitos de genótipos em cada ambiente conduz a um modelo para G, no qual:

$\sigma_{g_{jj}} = \sum_{r=1}^{k} \lambda_{jr}^2 + \psi_j$: variância genotípica no ambiente j;

$\sigma_{g_{jj'}} = \sum_{r=1}^{k} \lambda_{jr}\lambda_{j'r}$: covariância genotípica entre os ambientes j e j';

$\rho_{g_{jj'}} = \sum_{r=1}^{k} \lambda_{jr}\lambda_{j'r} / [(\sum_{r=1}^{k} \lambda_{jr}^2 + \psi_j)(\sum_{r=1}^{k} \lambda_{j'r}^2 + \psi_{j'})]^{1/2}$: correlação genotípica entre os ambientes j e j';

A equação para g_s tem a forma de uma regressão aleatória em k covariáveis ambientais $\lambda_1....\lambda_k$ nas quais todas as regressões passam pela origem. No entanto, pode ser mais apropriado permitir um intercepto separado (não zero) para cada genótipo. Isto é equivalente a um modelo com efeitos principais de genótipos, g, e um modelo fator analítico de ordem k para a interação g x e. Então, a expressão para g_s torna-se:

$$\begin{aligned} g_s &= (1_s \otimes I_g)g + ge \\ &= (1_s \otimes I_g)g + (\Lambda \otimes I_g)\,f + \delta. \end{aligned}$$

O vetor g tem média zero e variância $\sigma_g^2 I$ ou $\sigma_g^2 A$ em que A é a matriz de correlação genética aditiva entre os materiais genéticos. O modelo pode ser escrito como

$$g_s = (\sigma_g 1_s \otimes I_g) f_0 + (\Lambda \otimes I_g) f + \delta$$
$$= (\Lambda_g \otimes I_g) f_g + \delta,$$

em que:

$$\Lambda_g^{s(k+1)} = [\sigma_g 1_s \ \Lambda];$$
$$f_0 = g / \sigma_g;$$
$$f_g^{'} = (f_0^{'} f^{'}).$$

Então o modelo com efeitos principais de genótipos e um modelo fator analítico de ordem k para as interações g x e é um caso especial de um modelo fator analítico de ordem (k +1) em cada ambiente, no qual o primeiro conjunto de carregamentos são restritos a serem iguais. O modelo FAMM sem efeitos principais de genótipos é análogo aos modelos AMMI sem o ajuste para esses efeitos principais de genótipos e/ou ambiente, os quais são denominados modelos multiplicativos alterados ou modificados (SHMM – Shifted Multiplicative Models, conforme Cornelius et al., 1992), nos quais a PCA atua nas combinações genótipo-ambiente e não nas interações g x e.

A característica que distingue as equações para g_s dos problemas padrão de regressão aleatória multivariada é que no presente caso, ambos, as covariáveis e os coeficientes de regressão, são desconhecidos e então devem ser estimados dos dados. O modelo é então multiplicativo de coeficientes genotípicos e ambientais (conhecidos como escores fatoriais e carregamentos, respectivamente). Aqui está a analogia com os modelos AMMI. Entretanto, uma diferença fundamental é que o modelo multiplicativo na equação para g_s acomoda efeitos aleatórios, enquanto o AMMI é um modelo de efeitos fixos.

Para o modelo $y = Xb + Z[(\Lambda \otimes I_g) f + \delta] + \kappa + \varepsilon$, as equações de modelo misto e a estrutura de variâncias para um modelo (FAMMS) fator analítico com erros espaciais ($\varepsilon = \xi$) são dadas por

$$\begin{bmatrix} \hat{b} \\ \tilde{g}_s \\ \tilde{\kappa} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z+G^{-1} & Z'R^{-1}W \\ W'R^{-1}X & W'R^{-1}Z & W'R^{-1}W+C^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ W'R^{-1}y \end{bmatrix}$$ em que:

$$\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_s \end{bmatrix}; \qquad \tilde{g}_s = \begin{bmatrix} \tilde{g}_1 \\ \vdots \\ \tilde{g}_s \end{bmatrix}; \qquad \tilde{\kappa} = \begin{bmatrix} \tilde{\kappa}_1 \\ \vdots \\ \tilde{\kappa}_s \end{bmatrix}$$

$$R^{-1} = R_O^{-1} \otimes H^{-1}; \qquad G^{-1} = G_O^{-1} \otimes A^{-1}; \qquad C^{-1} = C_O^{-1} \otimes I$$

$$R_O = \begin{bmatrix} \sigma^2_{\xi_1} & 0 \\ 0 & \sigma^2_{\xi_s} \end{bmatrix}; \qquad G_O = \begin{bmatrix} \sigma_{g_{11}} & \sigma_{g_{1s}} \\ \sigma_{g_{1s}} & \sigma_{g_{ss}} \end{bmatrix}; \qquad C_O = \begin{bmatrix} \sigma^2_{\kappa_1} & 0 \\ 0 & \sigma^2_{\kappa_s} \end{bmatrix}$$, em que:

$$R^{-1} = \begin{bmatrix} H_1\sigma^2_{\xi_1} & 0 \\ 0 & H_s\sigma^2_{\xi_s} \end{bmatrix}^{-1} \qquad H = \begin{bmatrix} H_1 & 0 \\ 0 & H_s \end{bmatrix}$$

b e κ: vetores de efeitos fixos e aleatórios de parcelas, respectivamente.

$H_1 = [\sum_{c_1}(\Phi_{c_1}) \otimes \sum_{r_1}(\Phi_{r_1})]$: matriz de correlação espacial para o ambiente 1;

$H_s = [\sum_{c_s}(\Phi_{c_s}) \otimes \sum_{r_s}(\Phi_{r_s})]$: matriz de correlação espacial para o ambiente s;

σ^2_{ξ}: variância associada aos erros espaciais correlacionados.

Nesse caso, os efeitos principais de genótipos são ajustados implicitamente em $\tilde{g}_s = [\tilde{g}_1 \ldots \tilde{g}_s]'$. O ajuste explícito dos efeitos principais de genótipos é conseguido pela inclusão de um outro vetor aleatório para esses efeitos principais nas equações de modelo misto. Depois disso, os efeitos $\tilde{g}_s$ nas equações de modelo misto representarão interações g x e.

Resolvendo as equações de modelo misto acima, obtém-se os BLUP´s dos efeitos genotípicos em ambientes individuais. Os BLUP´s dos escores fatoriais *f* dos genótipos podem ser obtidos a partir de $\tilde{g}_s$ por

$$\tilde{f}_s = \hat{\text{var}}(f)[Z\ (\hat{\Lambda} \otimes I_g)]'\hat{P}y$$
$$= [\hat{\Lambda}'(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1} \otimes I_g]\tilde{g}_s.$$

As estimativas são:

$\hat{\Lambda}$: matriz dos carregamentos estimados;

$\hat{\Psi}$: matriz das variâncias específicas estimadas.

Os BLUP´s dos resíduos das interações g x e podem ser obtidos por

$$\tilde{\delta} = [\hat{\Psi}(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1} \otimes I_g]\tilde{g}_s.$$

Pode ser visto que o modelo fator analítico requer cálculo dos parâmetros Λ e Ψ, os quais compõem a matriz de variância-covariância G_0, e podem ser estimados por REML (Patterson e Thompson, 1971) por meio do algoritmo de informação média (AI) de Gilmour, Thompson e Cullis (1985) e Johnson e Thompson (1995). Um algoritmo REML específico para modelos fator analíticos foi desenvolvido por Thompson et al. (2003). Detalhes sobre o algoritmo AI são apresentados no Capítulo 4 e também em Cullis et al. (2004).

Pelo modelo $y = Xb + Z[(\Lambda \otimes I_g)f + \delta] + \varepsilon$, os efeitos genotípicos preditos em um ambiente médio ($\tilde{g}_{\bar{s}}$) são dados pela fórmula:

$\tilde{g}_{\bar{s}} = \bar{\hat{b}} + [(\bar{\hat{\lambda}}_1\ \bar{\hat{\lambda}}_2 .....\bar{\hat{\lambda}}_k) \otimes I_g]\tilde{f}$. É importante perceber a diferença de notação entre $\tilde{g}_{\bar{s}}$ e $\tilde{g}_s$, dada pelo $\bar{s}$ em lugar de s .

As quantidades $\bar{\hat{\lambda}}_r$ e $\tilde{f}$ são as médias através dos ambientes dos carregamentos estimados para o r-ésimo fator e os escores fatoriais dos genótipos, respectivamente. Esta é uma predição em termos dos valores médios dos carregamentos. Pela definição de carregamentos, estes são predições de médias genotípicas para um ambiente que é médio em termos de possuir uma covariância média com todos os outros ambientes. A predição da performance global do genótipo é a mesma, independentemente da inclusão do efeito principal de genótipos no modelo. A questão da interpretação da inclusão dos efeitos principais de genótipos é importante. Estes não são efeitos

principais no sentido usual de medida da performance genotípica global, mas são meramente interceptos na regressão. Então, refletem a performance do genótipo em um ambiente que tem valores zero para os carregamentos. A inclusão dos efeitos principais de genótipos propicia resultados idênticos àqueles valores preditos para um ambiente médio ($\tilde{g}_{\bar{s}}$) (Smith, Cullis e Thompson, 2001b; Resende e Thompson, 2003; 2004).

Uma forma de obtenção da performance global dos genótipos exatamente nos mesmos ambientes em que foram avaliados, é por meio da composição de uma tabela de dupla entrada das médias genotípicas preditas para cada ambiente e então obter as médias marginais desses valores através dos ambientes para obter a média global dos genótipos. Essas médias preditas são também dadas pela fórmula:

$$\tilde{g}_{\bar{s}m} = \bar{\hat{b}} + [(\bar{\hat{\lambda}}_1 \bar{\hat{\lambda}}_2 ..... \bar{\hat{\lambda}}_k) \otimes I_g]\tilde{f} + \bar{\hat{\delta}}$$

Esta fórmula difere de $\tilde{g}_{\bar{s}}$ somente pelo acréscimo da média dos efeitos não explicados de g x e, os quais se referem à falta de ajuste da análise de fatores. Essa performance global é um bom preditor somente se a correlação da performance genotípica entre ambientes é alta e os ambientes de plantio comercial forem os mesmos usados na avaliação.

Restrição e Rotação nos Fatores e Interpretação dos Carregamentos Ambientais e dos Escores Fatoriais

Quando o número k de fatores é maior do que 1, restrições devem ser impostas nos parâmetros fator analíticos, visando garantir identificabilidade. Isto acontece porque a distribuição de $(\Lambda \otimes I_g) f$ é singular. Pode ser demonstrado que k(k-1)/2 restrições independentes devem ser impostas nos elementos de Λ. De acordo com Mardia, Kent e Bibby (1988), o modelo fator analítico não é único sob rotação, então restrições devem ser escolhidas para garantir unicidade. Um grupo de restrições que preenche esse requisito refere-se a forçar todos os k(k-1)/2 elementos da parte triangular superior de Λ a ser zero, isto é, $\lambda_{jr} = 0$ para $j < r = 2...k$ (Jennrich e Schluchter, 1986). A implicação das restrições é que o número de parâmetros de variância no modelo fator analítico com k termos é dado por pk+p-k(k-1)/2 (Smith, Cullis e Thompson, 2001b).

A não unicidade de Λ quando k > 1 introduz ambigüidade na interpretação dos carregamentos ambientais e escores genotípicos. A forma restringida de Λ é meramente para facilidade computacional e não tem qualquer base biológica. Então, a rotação dos carregamentos é requerida para a geração de resultados com significado. Lawley e Maxwell (1971) descrevem várias rotações úteis. Para análise de dados de múltiplos experimentos a rotação requerida é $\Lambda^* = \Lambda T$, em que T é uma matriz ortogonal. De acordo com Johnson e Wichern (1988), os eixos podem então ser rotacionados em um certo ângulo ϕ e os carregamentos rotacionados podem ser dados por $\Lambda^* = \Lambda T$, com $T = \begin{bmatrix} \cos\phi & sen\,\phi \\ -sen\,\phi & \cos\phi \end{bmatrix}$.

Os carregamentos dos modelos fator analíticos são úteis no agrupamento de ambientes em termos de correlações genéticas. A dispersão gráfica dos carregamentos de um modelo com k > 1 pode ser muito informativo a esse respeito.

Na análise de fatores, o interesse principal está centrado nos parâmetros do modelo fatorial. Entretanto, os valores preditos dos fatores comuns, ou seja, os escores fatoriais são muito úteis na análise de agrupamento. Além da sua utilidade na predição de médias genotípicas, os escores fatoriais podem também ser plotados para os fatores 1 e 2, por exemplo, permitindo inferência sobre o agrupamento de genótipos com base em sua similaridade.

Comparação de Modelos e Procedimentos de Ajuste – Seleção de Modelos FAMM

Na procura por modelos parcimoniosos, os modelos FAMM de várias ordens k podem ser formalmente testados e adequados, já que são ajustados via abordagem de modelos mistos. O modelo com k fatores, denominado FAk, é hierárquico dentro do modelo com k + 1 fatores. Modelos incluindo os efeitos principais de genótipos (g) são intermediários entre os modelos fator analíticos de ordem k (FAk) e aqueles de ordem FAk+1. O modelo FA1 + g é intermediário entre os modelos FA1 e FA2. Testes de razão de verossimilhança residual (REMLLRT) podem ser usados para comparar tais modelos.

Outras abordagens para testar o ajuste de modelos fator analíticos envolvem comparações com matrizes de covariância não estruturadas (Mardia, Kent e Bibby,1988), as quais são de difícil ou impossível obtenção com grande número de ambientes.

Aplicação

Resultados referentes à avaliação de vários modelos aplicados a dados de genótipos de eucalipto avaliados em seis ambientes são apresentados na Tabela 31, conforme obtidos de Resende & Thompson (2003; 2004). Empregou-se os softwares ASREML (Gilmour et al., 2002) e Genstat (Thompson e Welham, 2003).

Tabela 31. Logaritmo da função de verossimilhança residual (Log L) maximizada e teste da razão de verossimilhança em relação ao modelo imediatamente precedente na tabela (REMLRT) para seqüências de modelos ajustados a dados de genótipos de eucalipto avaliados em seis ambientes.

Modelo para G	Log L	REMLRT	Número de Parâmetros de Variância em G	Número total de Parâmetros de Variância
1.Uniforme para g em cada ambiente	-151100	-	1	3
2.Uniforme para g	-149228	-	1	3
3.Uniforme para g + g x e	-147892	2672	2	4
4.FA1, var. homog.	-147619	546	12	14
5.FA2, var. homog.	-147562	114	17	19
6.Multiv.var. homog.	-147556	12	21	23
7.FA1, var. heterog.	-146381	-	12	19
8.FA1+g, var.heterog.	-146381	0	13	20
9.FA2, var. heterog.	-146325	112	17	24
10.Multiv. var. heterog.	-146318	14	21	28

A primeira parte da Tabela 31 contém somente modelos (1 a 6) ajustados com a suposição de homogeneidade de variância residual. O modelo 1 ajustou efeitos de tratamentos ou genótipos em cada ambiente e considerou uma variância residual comum para todos os ambientes. O modelo 2 ajustou efeitos de tratamentos ou genótipos para um ambiente médio e considerou uma variância residual comum para todos os ambientes. O modelo 3 ajustou efeitos principais de tratamentos ou genótipos mais os efeitos de interação g x e e considerou uma variância residual comum para todos os ambientes. O modelo 4 ajustou uma estrutura fator analítica de ordem 1 para os efeitos de tratamentos ou genótipos em cada ambiente e considerou uma variância residual comum para todos os ambientes. O modelo 5 ajustou uma estrutura fator analítica de ordem 2 para os efeitos de tratamentos ou genótipos em cada ambiente e considerou uma variância residual comum para todos os ambientes. O modelo 6 ajustou um modelo multivariado completo com matriz de variância e covariância não estruturada para os efeitos de genótipos e considerou uma variância residual comum para todos os ambientes.

A segunda parte da mesma Tabela 31 contém somente modelos (7 a 10), permitindo e considerando a heterogeneidade de variância residual. Os modelos 7 e 9 ajustaram uma estrutura fator analítica de ordem 1 e 2 para os efeitos de genótipos em cada ambiente, respectivamente. O modelo 8 ajustou uma estrutura fator analítica de ordem 1 para os efeitos de interação g x e mais os efeitos principais de genótipos. O modelo 10 ajustou um modelo multivariado completo com matriz de variância e covariância não estruturada para os efeitos de genótipos.

Contrastando as duas partes da Tabela 31 em termos do Log L, pode-se observar que os modelos que permitem considerar a heterogeneidade de variância residual são melhores do que aqueles que admitiram homogeneidade de variância residual. Isto mostra a superioridade dos modelos FAMM sobre os modelos AMMI, os quais não consideram a heterogeneidade de variância residual. A suposição de variâncias residuais comuns é implícita na abordagem AMMI. Até mesmo o modelo multivariado completo (6) para G_0 (21 parâmetros) com variância homogênea é pior do que o modelo FA1 (7) para G_0 (12 parâmetros) com variâncias residuais heterogêneas. Isto confirma a grande importância de considerar a heterogeneidade de variância residual na análise de múltiplos experimentos. E isto pode ser feito quando se adota a abordagem de modelos mistos. Dessa forma, o uso de modelos fator analíticos embutidos na metodologia de modelos mistos é muito vantajoso.

Outra importante característica dos modelos FAMM é obtenção de modelos parcimoniosos em relação à abordagem multivariada completa com matriz de variância e covariância não estruturada. Essa abordagem é proibitiva com número de ambientes maior que 3 ou 5, gerando modelos superparametrizados e de difícil convergência. Resultados da Tabela 31 revelam que o modelo FAMM com dois fatores (FA2) é praticamente equivalente (REMLRT de 14 e 12 com 4 graus de liberdade, p valor > .01) ao modelo multivariado completo em ambas as situações, com e sem permissão para heterogeneidade de variâncias. Então, na prática modelos com quatro parâmetros a menos podem ser usados. É importante mencionar que os modelos FAMM convergiram sem a necessidade de restrição na matriz G_0.

Um modelo incluindo os efeitos principais de genótipos (g) é intermediário entre modelos fator analíticos de ordem k (FAk) e de ordem FAk+1, porque ele é um modelo FAk+1 com restrições. O modelo FA1 + g é intermediário aos modelos FA1 e FA2. No presente conjunto de dados, os modelos FA1 e FA1 + g foram equivalentes, produzindo o mesmo Log L. De fato, a estimativa do componente de variância para os efeitos genotípicos foi estimado como zero. O papel dos efeitos principais de genótipos em um modelo FA é puramente em termos de busca de uma estrutura de variância parcimoniosa entre um dado modelo FAk e outro modelo FAk+1. A abordagem para a predição das médias genotípicas globais através dos ambientes é a mesma independentemente da inclusão dos efeitos principais de genótipos (Smith, Cullis e Thompson, 2001b). Em um contexto fator analítico, o modelo sem os efeitos principais de genótipos é equivalente a um modelo para os efeitos genotípicos em cada local.

No geral, o melhor modelo parcimonioso foi o FA2 com variâncias residuais heterogêneas (modelo 9 na Tabela 31). Os resultados referentes aos carregamentos, variâncias comuns, variâncias específicas e variâncias do erro associados a esse modelo são apresentados na Tabela 32.

Tabela 32. Carregamentos estimados (na escala de correlações), variâncias comuns (comunalidades), variâncias específicas e variâncias do erro para o modelo FA2 ajustado a dados de eucalipto.

	Carregamentos Originais e (Rotacionados)		**Variância Comum (%)**	**Variância Específica(%)**	**Variância do Erro**
Locais	**Fator 1**	**Fator 2**			
1.L1	0.845 (0.433)	0.498 (0.880)	0.962	0.038	20.0422
2.L2	0.791 (0.443)	0.398 (0.767)	0.784	0.216	20.5270
3.L3	0.837 (0.450)	0.454 (0.839)	0.907	0.093	22.6041
4.L4	0.907 (0.596)	0.295 (0.745)	0.910	0.090	44.5751
5.L5	0.979 (0.761)	0.104 (0.624)	0.969	0.031	38.0380
6.L6	0.904 (0.837)	-0.149 (0.372)	0.839	0.161	28.9856
Autovalores	4.639	0.710			
Proporção de Variação Explicada Acumulada	0.773	0.892			

Pode ser visto que o modelo FA2 explicou uma grande quantidade (quase 90 %) da variação genotípica total. O primeiro fator explicou 77.3 % da variação total e o segundo fator somou mais 11.9 %. As variâncias específicas (em porcentagem do total) foram baixas, exceto para os ambientes 2 e 6, em que equivaleram a 22 % e 16 %, respectivamente. Os altos valores de variância comum (ou comunalidade) mostram que os dois fatores explicaram uma grande porcentagem da variância de cada ambiente e que o modelo FA2 ajustou-se muito bem ao conjunto de dados (Tabela 32).

A matriz de variância-covariância e de correlações genotípicas através dos ambientes (obtidas por $\Lambda\Lambda' + \Psi$ do modelo FA2 na escala de correlações) são apresentadas na Tabela 33.

Tabela 33. Matriz de covariância\variância\correlação genotípica associada ao modelo FA2, aplicado ao conjunto de dados de eucalipto em seis locais (L).

	L1	**L2**	**L3**	**L4**	**L5**	**L6**
L1	6.312	0.867	0.933	0.914	0.879	0.689
L2	6.964	10.225	0.843	0.835	0.812	0.655
L3	7.375	8.481	9.905	0.893	0.867	0.689
L4	8.132	9.463	9.959	12.555	0.919	0.776
L5	6.566	7.754	8.108	9.682	8.837	0.869
L6	5.135	6.207	6.425	8.148	7.659	8.784

Pode-se observar que existe heterogeneidade entre variâncias específicas referentes aos vários ambientes (diagonal da Tabela 33). Isto justifica o uso de modelos com variâncias específicas heterogêneas. Piepho (1997, 1998) propôs o uso de modelos fator analíticos com variâncias específicas comuns para todos os ambientes. Entretanto, Smith, Cullis e Thompson (2001b) notaram que modelos com variâncias específicas heterogêneas foram significativamente melhores. Pode-de notar, também, a existência de heterogeneidade de covariâncias entre as várias combinações de locais. Essas covariâncias representam a variância genotípica livre dos efeitos da interação entre cada dois sítios. Essa heterogeneidade explica o melhor ajuste do modelo FAk e modelo multivariado sobre o modelo 3, o qual inclui g + g x e. Quando existem somente dois ambientes, o modelo bivariado e o modelo 3 tendem a propiciar o mesmo ajuste desde que se corrija a heterogeneidade de variância (ver os tópicos 5 e 6).

Os resultados sobre correlações revelaram que os quatro primeiros ambientes apresentam menores correlações com o ambiente 6, o qual tem correlação mais alta com o ambiente 5 (Tabela 33). Pode ser observado que a análise de fatores coloca maior ênfase nos ambientes 5 e 6 no fator 1 (carregamentos rotacionados maiores que 0.76) e maior ênfase nos locais 1, 2, 3 e 4 no fator 2 (carregamentos rotacionados maiores que 0.74) (Tabela 32). Essa é a lógica da análise de fatores: separar grupos de caracteres ou ambientes com altas correlações entre eles em cada grupo e então colocar maiores pesos ou cargas em caracteres de um grupo em um fator (fator 1) e maiores pesos em caracteres de outro grupo em outro fator (fator 2). Fazendo a dispersão gráfica do primeiro grupo de carregamentos contra o segundo, isso mostrará o agrupamento de ambientes: L1, L2, L3 e L4

muito próximos em um grupo e L5 e L6 em um segundo grupo, sendo que L5 também não se distanciará muito do primeiro grupo. Outra vantagem dos modelos FAMM sobre os modelos AMMI é que os primeiros propiciam uma estimação da completa estrutura de correlação genotípica, facilitando a tomada de decisões práticas com base nessas informações.

Os modelos FAMM e AMMI são também úteis no agrupamento de ambientes com base em suas similaridades em termos de correlações genéticas. Isto pode ser feito, por exemplo, usando *biplots*. A completa estrutura de correlação propiciada pelos modelos FAMM pode ser submetida a métodos multivariados de agrupamento. Tais métodos tradicionalmente operam em correlações estimadas para pares de locais por meio do procedimento ANOVA para dados balanceados. Os modelos FAMM usam a informação de todos os ambientes simultaneamente para fornecer as correlações para pares de ambientes. Assim, propiciam estimativas mais precisas.

5 COMPARAÇÃO ENTRE ESTRUTURAS DE COVARIÂNCIA NA ANÁLISE DE MÚLTIPLOS EXPERIMENTOS

Os experimentos com genótipos repetidos em vários locais (e também os experimentos envolvendo várias gerações de avaliação) caracterizam-se pela presença de correlação em nível dos efeitos de tratamentos ou genótipos através dos locais. Os demais efeitos (por exemplo, blocos e resíduo) são não correlacionados através dos locais. A modelagem completa e ótima dos efeitos de tratamentos é propiciada pelo modelo multivariado, com matriz de covariância não estruturada (UN), o qual trata os vários locais como se fossem diferentes caracteres, conforme descrito e aplicado nesse contexto por Resende et al. (1999) e também recomendado recentemente por Piepho e Mohring (2006). Nesse caso, contempla-se tanto a heterogeneidade de variâncias quanto a de covariâncias entre locais, as quais conduzem também à heterogeneidade de correlações entre pares de locais. No entanto, essa modelagem é a mais complexa possível e, com grande número de locais, é proibitiva devido à necessidade de estimação de um grande número de parâmetros e à dificuldade de convergência na análise.

A modelagem mais simples e parcimoniosa possível baseia-se na estrutura de simetria composta (CS), a qual assume tanto homogeneidade de variâncias quanto de covariâncias entre locais. Essa abordagem é desejável porque depende do menor número possível de estimativas de parâmetros. No entanto, pode ser ineficiente no caso de grande heterogeneidade de variância e de covariância entre locais. A modelagem via estrutura CS é exatamente igual à abordagem de um modelo univariado misto com efeito de genótipos e de interação (ou combinação) genótipo x ambiente, conforme descrito por Resende (2002;2004) e Piepho e Mohring (2005) e apresentado no Capítulo 11. Para uma aplicação segura dessa estrutura CS, recomenda-se, em caso de presença de grande heterogeneidade de variâncias, a correção prévia dos dados de cada local por meio da multiplicação dos mesmos pela razão hi/hm, em que hi refere-se à raiz quadrada herdabilidade no local i, e hm refere-se à raiz quadrada da média das herdabilidades nos vários ambientes, conforme Resende (2004) e tópico 6, a seguir. Esse procedimento considera simultaneamente a heterogeneidade de variância genética e residual. Procedendo-se dessa forma, a aplicação da estrutura CS fornece resultados semelhantes àqueles que se obtém ao se aplicar a estrutura de simetria composta com variâncias heterogêneas (CSH).

No caso de análises conjuntas de locais dois a dois e com presença de homogeneidade de variâncias (ou corrigindo a heterogeneidade), a aplicação das estruturas UN e CS produzem resultados semelhantes. Ainda no caso de análises de locais dois a dois, com heterogeneidade de variâncias, as estruturas UN e CSH são idênticas. Com análises envolvendo mais que dois locais e com presença de homogeneidade de variâncias e covariância (ou corrigindo a heterogeneidade), as estruturas UN e CSH produzem resultados semelhantes. Em resumo, com correção para a heterogeneidade de variâncias e em presença de correlações genéticas não muito discrepantes entre pares de locais, não há necessidade de se usar o modelo multivariado ou estrutura UN.

Com correlações genéticas muito discrepantes entre pares de locais e grande número de locais, outras estruturas intermediárias entre UN e CS podem ser usadas. Dentre essas, citam-se: fator analítica multiplicativa sob modelos mistos (FAMM), que é análoga ao modelo AMMI, porém considera genótipos como efeitos aleatórios (Piepho, 1998; Smith et al., 2001; Resende e Thompson, 2003; 2004); componentes principais genéticos sob modelos mistos (Meyer e Kirkpatrick, 2005); autoregressiva com variâncias heterogêneas (ARH); antedependência estruturada (SAD);

Toeplitz ou de correlação bandada; modelos de regressão aleatória; ajuste de *splines*. Essas estruturas são usadas para modelagem de matrizes de covariância e de correlação em vários contextos, sejam eles espaciais (entre locais e dentro de locais) ou temporais (medidas repetidas no tempo). Entretanto, no caso de ensaios multi-locais com vegetais, apenas a estrutura FAMM parece adequada, permitindo considerar, sob modelos parcimoniosos, a presença de heterogeneidade das correlações entre locais. Os demais modelos citados são mais adequados a estudos de medidas repetidas no tempo, em que as correlações seguem alguns padrões em função das distâncias entre as medidas consideradas, em geral havendo decréscimo das correlações em função do aumento dessas distâncias. Esse não é o caso das correlações envolvendo diferentes ambientes. Esses métodos são melhor comentados no Capítulo 9.

6 CORREÇÃO PARA HETEROGENEIDADE DE VARIÂNCIAS

Uma das correções para heterogeneidade de variância mais usadas na literatura refere-se à multiplicação dos dados de cada ambiente pela razão desvio padrão fenotípico médio (Sm) para todos os ambientes / desvio padrão fenotípico no ambiente i (Si), conforme Visscher et al. (1992). Esta correção considera apenas a heterogeneidade de variância fenotípica. Outras correções usam os inversos dos desvios padrões genético (Sg) ou ambiental (Se) em lugar do fenotípico (Sf). Padronizar pelo inverso do desvio padrão genético, conforme realizado por Dutkowski et al. (2006) e McRae et al. (2004), parece um contra-senso. Isto porque são penalizados aqueles ambientes com maior expressão de variação genética e, possivelmente, com maior herdabilidade. Padronizar pelo inverso de Sf também pode conduzir a isso pois Sf contém uma função de Sg. Visando contornar isso, e ao mesmo tempo padronizar para uma escala média, Resende (2004) propôs o seguinte fator de correção para multiplicar os dados $Fc = (Sf_m/Sf_i)\ (Sg_i/Sg_m) = h_i/h_m$, em que os índices i e m referem-se aos ambientes específicos i e média dos ambientes, respectivamente. Essa expressão penaliza os ambientes com alta variação fenotípica (divisão por Sf_i) desde que não apresentem alta variação genética (multiplicação por Sg_i, como proporção da variação genética média Sg_m). Em outras palavras, penaliza ambientes com alta variação ambiental. Dito de outra forma, beneficia ambientes com maior raiz quadrada da herdabilidade (h_i) em relação à raiz quadrada das médias de

herdabilidade em todos os ambientes (h_m), ou seja, beneficia ambientes com maior expressão de variabilidade ou que contemplam materiais genéticos mais variáveis. Isso é coerente e desejável. Além disso, considera as herdabilidades em ambientes individuais à semelhança da análise multivariada.

Essa correção Fc é similar à correção pelo inverso do desvio padrão ambiental Se. Isto pode ser visto, pela igualdade $Se_m/Se = (Sf_m/Sf_i)\ (1-h^2_m)^{½}/\ (1-h^2_i)^{½}$. O primeiro termo dessa expressão é igual ao primeiro termo da expressão Fc e o segundo termo $[(1-h^2_m)^{½}/\ (1-h^2_i)^{½}]$ é similar ao segundo termo (Sg_i/Sg_m) de Fc, no sentido de que quanto maior Sg_i , maior h^2_i e menor $(1-h^2_i)$, ou seja, maior é o peso em ambas as correções. Recentemente (Costa e Silva et al., 2005) concluíram que a correção por Se foi adequada para aproximar um modelo multivariado e mostrou-se superior à correção por Sf. Também Thompson (1996) e Smith et al. (2001) ponderam os dados de experimentos individuais por Se_m/Se antes da realização de análises conjuntas na Inglaterra e Austrália, respectivamente. O que foi exposto até aqui corrobora que as correções por Fc e Se_m/Se relatadas nesse texto são adequadas.

A idéia de aplicar a razão entre as raízes quadradas da herdabilidade no ambiente i (hi) e da média das herdabilidades em todos os ambientes (hm) visa considerar tanto a heterogeneidade de variância genética quanto residual. Considerando que o preditor BLUP aplicado na análise de dados de todos os ambientes simultaneamente, pondera as informações por uma herdabilidade média válida para todas os locais, verifica-se que o valor esperado da ponderação final dos dados de cada ambiente, realizada pela correção, seguida pela aplicação do BLUP, é dada aproximadamente por $(hi/hm) * h^2m = hi\ hm$. Verifica-se que essa ponderação depende simultaneamente da herdabilidade do ambiente alvo da seleção (no caso, o ambiente médio de todas as safras) e da confiabilidade dos dados de cada ambiente, dada por uma função (hi) da herdabilidade de cada ambiente. Quanto menor a herdabilidade de um ambiente, menor é o peso dado à informação desse ambiente. Quanto maior a herdabilidade em um ambiente, maiores são a expressão de variabilidade genética e/ou a precisão experimental nesse ambiente. Então, maior é o peso dado à informação desse ambiente Isto é coerente e desejável na prática.

Isto também foi comprovado via um estudo de validação, visando comparar a eficiência da aproximação univariada em reproduzir resultados da estrutura multivariada. Usando dados reais da

avaliação de 50 genótipos de eucalipto em seis locais e um modelo multivariado (matriz de covariância não estruturada, portanto, levando em conta completamente a heterogeneidade de variâncias) considerando todos os dados simultaneamente, os efeitos genotípicos foram preditos para cada ambiente e foram obtidos os efeitos genotípicos médios para o ambiente médio. Efeitos genotípicos para o ambiente médio foram preditos também pela média dos efeitos genéticos preditos por modelos univariados do tipo g + ge,usando as transformações (Sf_m/Sf_i), hi/hm e (Sg_m/Sg_i). Os resultados obtidos pelo modelo multivariado completo foram considerados como paramétricos ou verdadeiros e os erros de predição pelos demais métodos univariados foram calculados em relação aos resultados do modelo multivariado. Tais erros de predição foram apresentados em porcentagem do valor paramétrico, sendo dados pela expressão $erro = \left|(\hat{\theta} - \theta_0)/\theta_0\right| * 100$, em que $\hat{\theta}$ e θ_0 referem-se aos efeitos genotípicos preditos e paramétricos (modelo multivariado), respectivamente. Os resultados são apresentados na Tabela 34.

Tabela 34. Efeitos genotípicos preditos pelos modelos multivariado (Multiv.), univariado sem correção ou transformação prévia dos dados (Univ. NT), univariado com correção prévia dos dados dada por (sf_m/sf_i) (Univ. Sf), univariado com correção prévia dos dados dada por hi/hm (Univ. hi/hm) e univariado com correção prévia dos dados dada por (sg_m/sg_i) (Univ. Sg), bem como erros de predição desses efeitos, em porcentagem (erro %).

Genótipo	Multiv.	Erro (%)	Univ. NT	Erro (%)	Univ. Sf	Erro (%)	Univ: hi/hm	Erro (%)	Univ: Sg	Erro (%)
1	2.07	0.00	2.08	0.74	2.02	2.11	2.09	0.83	1.98	4.04
2	1.68	0.00	1.68	0.08	1.64	2.32	1.69	0.62	1.61	4.32
3	-3.77	0.00	-3.80	0.86	-3.80	0.97	-3.80	0.88	-3.86	2.58
4	0.50	0.00	0.54	8.23	0.49	1.76	0.47	5.61	0.52	3.83
5	1.38	0.00	1.42	2.66	1.45	4.47	1.40	1.33	1.52	9.52
6	0.54	0.00	0.61	**13.39**	0.50	7.29	0.51	5.25	0.50	7.93
7	-0.42	0.00	-0.44	5.99	-0.55	**30.96**	-0.42	1.52	-0.69	**66.51**
8	-1.35	0.00	-1.33	2.08	-1.33	1.86	-1.37	1.38	-1.30	4.11
9	-4.49	0.00	-4.41	1.70	-4.50	0.25	-4.42	1.57	-4.66	3.89
10	-0.92	0.00	-0.98	6.93	-0.90	1.49	-0.91	0.89	-0.91	0.31
11	-4.02	0.00	-4.04	0.29	-3.98	1.19	-4.03	0.18	-3.97	1.42
12	3.87	0.00	3.91	0.93	3.85	0.55	3.85	0.46	3.90	0.74
13	0.95	0.00	1.11	**17.14**	1.01	6.36	0.96	0.83	1.08	**14.16**
14	0.66	0.00	0.71	7.12	0.58	**13.51**	0.69	3.89	0.44	**34.22**
15	-0.39	0.00	-0.39	1.28	-0.35	8.28	-0.43	**10.57**	-0.27	**29.46**
16	-1.09	0.00	-1.17	6.58	-1.15	5.14	-1.06	2.90	-1.28	**17.13**
17	4.06	0.00	4.09	0.79	4.00	1.56	4.07	0.17	3.97	2.17
18	-1.83	0.00	-1.85	1.54	-1.85	1.51	-1.82	0.33	-1.93	5.80
19	-6.31	0.00	-6.31	0.04	-6.33	0.28	-6.29	0.31	-6.47	2.51
20	0.77	0.00	0.75	2.54	0.78	1.79	0.80	4.29	0.75	2.04
21	-0.53	0.00	-0.54	2.32	-0.49	7.57	-0.55	4.09	-0.42	**21.60**
22	-2.21	0.00	-2.16	2.09	-2.28	3.07	-2.22	0.27	-2.39	8.01
23	4.43	0.00	4.44	0.33	4.40	0.77	4.46	0.61	4.39	0.92
24	-0.91	0.00	-0.90	1.33	-1.01	**11.30**	-0.91	0.65	-1.13	**24.62**
25	-0.52	0.00	-0.50	4.18	-0.46	**12.02**	-0.49	7.24	-0.45	**15.04**
26	3.62	0.00	3.65	0.90	3.61	0.25	3.64	0.49	3.62	0.06
27	2.87	0.00	2.94	2.20	2.77	3.63	2.84	1.29	2.73	4.83
28	0.81	0.00	0.81	0.35	0.69	**14.28**	0.84	3.71	0.54	**33.56**

segue

conclusão Tabela 34

Genótipo	Multiv.	Erro (%)	Univ: NT	Erro (%)	Univ: Sf	Erro (%)	Univ: hi/hm	Erro (%)	Univ: Sg	Erro (%)
29	-0.82	0.00	-0.80	1.82	-0.82	0.78	-0.82	0.76	-0.83	2.33
30	-4.38	0.00	-4.48	2.47	-4.50	2.88	-4.42	1.06	-4.68	6.96
31	1.63	0.00	1.65	1.12	1.69	3.63	1.64	0.32	1.77	8.55
32	-8.04	0.00	-7.96	0.98	-8.03	0.14	-8.06	0.27	-8.11	0.91
33	-2.85	0.00	-2.87	0.56	-2.94	3.13	-2.95	3.26	-2.96	3.56
34	-6.22	0.00	-6.16	0.93	-6.29	1.16	-6.24	0.46	-6.42	3.23
35	-1.14	0.00	-1.11	2.37	-1.13	0.31	-1.17	2.80	-1.10	3.29
36	-0.36	0.00	-0.30	**15.30**	-0.35	1.80	-0.39	9.02	-0.29	**17.59**
37	1.85	0.00	1.81	1.75	1.89	2.46	1.91	3.69	1.89	2.41
38	-2.40	0.00	-2.54	5.91	-2.20	8.18	-2.38	0.78	-2.04	**15.03**
39	2.93	0.00	2.90	1.07	2.98	1.65	2.94	0.26	3.09	5.33
40	1.84	0.00	1.92	4.52	1.89	2.85	1.82	1.22	2.01	9.06
41	2.76	0.00	2.60	5.56	2.91	5.42	2.81	2.08	3.05	**10.60**
42	2.25	0.00	2.32	3.04	2.31	2.74	2.25	0.23	2.41	7.17
43	4.88	0.00	4.80	1.68	5.09	4.31	4.93	0.99	5.35	9.76
44	0.37	0.00	0.38	3.43	0.49	**32.77**	0.36	3.12	0.66	**78.02**
45	2.76	0.00	2.69	2.48	2.84	3.25	2.80	1.59	2.94	6.77
46	3.47	0.00	3.39	2.44	3.36	3.39	3.44	1.13	3.32	4.41
47	2.25	0.00	2.40	6.51	2.20	2.07	2.28	1.11	2.15	4.45
48	-1.57	0.00	-1.68	7.02	-1.47	6.72	-1.61	2.15	-1.32	**16.17**
49	-7.58	0.00	-7.58	0.04	-7.65	0.98	-7.66	0.99	-7.75	2.22
50	8.90	0.00	8.75	1.62	8.93	0.40	8.95	0.54	9.04	1.58
		0.00		**3.34**		**4.75**		**2.00**		**10.89**

É importante enfatizar que os erros estão apresentados em módulo ou valor absoluto. Na apresentação, os efeitos genotípicos preditos foram representados com apenas duas decimais. Assim, se calculados a partir da presente tabela, os erros diferirão ligeiramente dos que estão apresentados na própria tabela. Verifica-se que o modelo univariado com transformação hi/hm mostrou-se eficiente, conduzindo a um erro médio de predição de apenas 2 % em relação ao modelo multivariado, o qual considera perfeitamente a heterogeneidade de variâncias. Tal transformação

apresentou apenas um erro superior a 10 % (valor em negrito na Tabela 34), dentre os 50 genótipos. O modelo univariado com transformação (Sf_m/Sf_i) conduziu a um erro médio de 4.75 % e seis erros superiores a 10 %, confirmando a superioridade do método que usa hi/hm.

O modelo univariado do tipo g + ge sem qualquer correção apresentou melhor resultado do que o modelo univariado com transformação (Sf_m/Sf_i), produzindo um erro médio de 3.34 % e três erros superiores a 10 %. Isto confirma que a padronização pelo inverso do desvio padrão fenotípico não é adequada, sobretudo porque pode penalizar ambientes com maior variação genética. O modelo univariado com transformação (Sg_m/Sg_i) conduziu a um erro médio de 10.89 % e 14 erros superiores a 10 %, confirmando que essa é uma correção totalmente inadequada.

Essa validação foi aplicada a um conjunto de dados com correlação genotípica média através dos seis locais igual a 0.75 e heterogeneidade de variância no limite de significância pelo teste de Hartley. Assim, os resultados são válidos para essas condições. Entretanto, tais resultados evidenciam claramente a tendência esperada quando se usam esses tipos de correção ou padronização de dados.

CAPÍTULO 9

ANÁLISE ESTATÍSTICA DE MEDIDAS REPETIDAS

1 MÉTODOS DE ANÁLISE DE DADOS LONGITUDINAIS OU MEDIDAS REPETIDAS

Caracteres avaliados repetidas vezes no decorrer da vida dos organismos (plantas ou animais) são denominados infinitamente dimensionais, no sentido de que, em cada unidade de tempo ou idade, o caráter pode ser avaliado, gerando um conjunto multidimensional de dados. O interesse na análise desse tipo de dados geralmente reside na predição de valores dos indivíduos para determinado ponto no tempo ou através de todos os pontos e também na identificação de uma parcimoniosa estrutura de variância ao longo do tempo.

No caso de medidas repetidas em cada indivíduo (ou tratamento) ao longo do tempo, várias alternativas (Resende e Sturion, 2001; Resende, 2002a, p.522; Resende e Thompson, 2003; Gilmour, 2006) existem para a predição de efeitos e modelagem da estrutura de correlação entre as referidas medidas: (i) modelo univariado simplificado de repetibilidade, o qual assume que o caráter

é o mesmo (correlação genética igual a 1 através do tempo) de uma medição para outra, que as correlações fenotípicas (repetibilidades) são de iguais magnitudes entre todos os pares de idade e que as variâncias (genética e residual) são homogêneas; (ii) modelo univariado de repetibilidade mais interação genótipos x medições; (iii) modelo multivariado completo com matriz de covariância não estruturada, assumindo cada medida como um caráter diferente; (iv) modelo de regressão aleatória parcimonioso como aproximação do modelo multivariado; (v) ajuste de curva *spline* cúbica ou, alternativamente, *spline* tipo B, no intervalo de idades considerado; (vi) modelos processo caráter, tal como o modelo auto-regressivo com variâncias heterogêneas (ARH); (vii) modelos ante-dependência estruturados (SAD); (viii) modelo de correlação bandada (estrutura Toeplitz) com correlações específicas para cada intervalo entre medições; (ix) estrutura de simetria composta (CS); (x) estrutura de simetria composta com variâncias heterogêneas (CSH).

No melhoramento animal, esses dados multidimensionais associados a lactações, crescimento em peso, produção de lã e produção de ovos, geralmente são analisados via regressão aleatória, *splines*, modelos processo caráter ou SAD. Gilmour (2006) destacam os dois últimos como mais eficientes e relata que as propriedades matemáticas da regressão polinomial revelam que o ajuste de *splines* cúbicas é mais eficiente que o uso da técnica de regressão aleatória.

Em alguns casos, os modelos de repetibilidade são também suficientes. Quanto ao uso do modelo multivariado em lugar do modelo de repetibilidade, pode-se dizer que o modelo de repetibilidade é muito eficiente quando comparado ao multivariado, obtendo-se perdas de apenas 0 % a 5 % em eficiência, quando a correlação genética entre medições sucessivas é alta (acima de 0,80). Nestes casos, a correlação entre ordenamentos pelo modelo de repetibilidade e pelo modelo multivariado é muito próxima de 1 (Mrode, 2005). O grande benefício do modelo de repetibilidade comparado com o multivariado refere-se à menor demanda computacional e à necessidade de estimativas de poucos parâmetros genéticos.

Em bovinos de corte, suínos, ovinos de corte e aves de corte, muitas vezes os caracteres de interesse são pesos em algumas idades pré-estabelecidas e tais caracteres são tomados como caracteres distintos, os quais são posteriormente usados em índice de seleção. Nesse caso, os modelos para cada caráter contêm algumas combinações de efeitos fixos (grupo contemporâneo, por exemplo), alguma covariável (por exemplo, idade do animal) e os fatores aleatórios associados

ao valor genético aditivo direto e materno (que são correlacionados) e ambiente permanente materno. Em suínos e aves, os efeitos permanentes maternos são ajustados via os efeitos de família (leitegada), de forma confundida com os efeitos de capacidade específica de combinação e de ambiente comum à família. Em aves comerciais, os efeitos maternos podem ser desconsiderados.

Os caracteres de interesse em bovinos de leite, ovinos de lã e aves de postura não exibem efeitos maternos apreciáveis e os modelos não precisam considerar tais efeitos. Para esses caracteres, modelos com efeitos fixos, covariáveis e com os efeitos genéticos aditivos e de ambiente permanente da vaca são adequados em algumas situações. No entanto, o uso dos modelos de regressão aleatória tem sido comum recentemente e pode ser mais eficiente. Mrode et al. (2006) relatam que os modelos de repetibilidade e de regressão aleatória conduzem a seleções bastante diferenciadas. Esses últimos tendem a dar mais peso às informações individuais em detrimento das informações de família. Assim, os modelos de regressão aleatória tendem a selecionar mais animais jovens do que os modelos de repetibilidade. Em tais modelos de regressão aleatória, a estrutura regressiva é aplicada a todos os fatores aleatórios do modelo, ou seja, aos efeitos aditivos diretos e maternos e aos efeitos permanentes.

Em plantas perenes, o número de medições realizadas varia tipicamente de 3 a 6, pois um número maior de safras anuais compromete a eficiência dos programas de melhoramento por unidade de tempo. Com número de medições desta ordem, as técnicas de regressão aleatória e de *splines* tendem a não ser eficientes devido ao reduzido número de idades abrangido pelos dados. Tais técnicas são muito empregadas no melhoramento animal, em que indivíduos de diferentes idades (as medidas repetidas não ocorrem em intervalos fixos) são avaliados produzindo grande número de pontos em termos de idades. O modelo multivariado completo (não estruturado) tende a apresentar problemas de convergência em virtude das altas correlações geralmente verificadas entre medidas repetidas e do elevado número de parâmetros a ser estimado. A abordagem multivariada torna-se proibitiva quando o número de colheitas ou safras é elevado. Algumas vezes, mesmo com pequeno número de safras, as matrizes de covariância não estruturadas não são positivas definidas e conduzem a correlações maiores que 1. Medidas repetidas altamente correlacionadas aumentam o risco de obtenção de matrizes de covariância não positiva definidas e de não convergência no processo de estimação de componentes de variância. Também, quanto

maior a ordem da matriz de covariância, maior é a chance de que ela não seja positiva definida. O modelo de correlação bandada é mais parametrizado do que os modelos ARH, SAD, CS e CSH e só deve ser usado quando a correlação entre medidas distantes não puder ser representada adequadamente por uma função potência da correlação entre medidas adjacentes. Assim, as opções mais interessantes aos melhoristas de plantas perenes são os modelos de repetibilidade (quando as suposições são atendidas), repetibilidade + interação genótipos x medições, ARH, SAD, CS e CSH.

Para caracteres não associados a curvas de crescimento com a idade, tais quais produção de frutos, látex, massa foliar e açúcar, antes da aplicação do modelo de repetibilidade para todas as safras, é recomendável realizar análises individuais por safra, verificando a suposição de homogeneidade de variâncias genética e ambiental. Se esta suposição for rejeitada, recomenda-se a transformação h_i/h_{im} mencionada no Capítulo 8, a aplicação do modelo de repetibilidade + interação genótipos x medições e a obtenção do valor genotípico predito em cada safra via g + gm. Este procedimento aproxima bem o modelo multivariado que, teoricamente, é o mais eficiente. Os modelos ARH e SAD são parcimoniosos e também aproximam bem o modelo multivariado, sendo especialmente indicados para o caso em que as correlações diminuem gradativamente com o aumento da distância entre as idades.

Os caracteres de interesse no melhoramento de plantas perenes se expressam mais de uma vez em cada indivíduo, gerando dados longitudinais ou medidas repetidas. Dados desses caracteres apresentam estrutura correlacionada através do tempo, safras ou medições e alguns métodos de análise desses caracteres são especificadas na seqüência.

a) **Análise univariada considerando cada idade ou safra em separado** (estrutura de covariância não correlacionada).

Neste caso, as avaliações em diferentes estágios são consideradas como sendo caracteres diferentes e não correlacionados. Assim, os dados de cada idade ou safra são analisados separadamente.

Nesta situação, a estrutura de covariância genética entre quatro diferentes idades ou safras é dada por:

$$\sum_g = \begin{bmatrix} \sigma^2_{g_1} & 0 & 0 & 0 \\ 0 & \sigma^2_{g_2} & 0 & 0 \\ 0 & 0 & \sigma^2_{g_3} & 0 \\ 0 & 0 & 0 & \sigma^2_{g_4} \end{bmatrix}$$, em que a $\sigma^2_{g_i}$ é a variância genética na idade i.

Este modelo não é realista e deve ser evitado. Em algumas plantas perenes, uma só avaliação, às vezes, é suficiente. Por exemplo, em espécies florestais como o eucalipto, a seleção precoce é enfatizada, procurando-se uma idade mínima que torne a seleção eficiente. Em geral, a maioria dos programas conduzidos com esta espécie no Brasil utiliza eficientemente a seleção baseada em uma só avaliação aos três anos.

Em outras plantas perenes como as fruteiras, muitas vezes, é gerado apenas um dado por indivíduo, dado este referente à soma ou média de várias safras. A seleção baseada apenas nesse dado totalizado ou médio raramente será completamente eficiente, a menos que todos os indivíduos tenham apresentado produção em todas as safras. Por outro lado, a seleção baseada em médias por grupo de plantas e em várias safras, como a seleção de genitores baseada na progênie, a seleção de clones em testes clonais e a seleção de híbridos ou famílias híbridas, tende a ser menos prejudicada pelo uso de um único dado médio.

b) **Análise pelo modelo de repetibilidade** (estrutura de covariância genética completamente correlacionada).

Nesta situação, considera-se que o caráter é o mesmo através das várias safras. Isto implica assumir que a correlação genética entre qualquer par de safras equivale a 1, que a variância fenotípica é a mesma para todas as safras e que a correlação ambiental permanente é a mesma

para todos os pares de safras. Essas premissas são atendidas aproximadamente para algumas fruteiras, mas não para todas. Todo o conjunto de dados é analisado simultaneamente e a estrutura de covariância genética entre quatro diferentes safras é dada por:

$$\sum_g = \sigma_g^2 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$, em que σ_g^2 é a variância genética.

A estrutura geral de covariância fenotípica entre quatro diferentes safras é dada por:

$$\sum = \begin{bmatrix} k & a & a & a \\ a & k & a & a \\ a & a & k & a \\ a & a & a & k \end{bmatrix},$$

em que *k* representa os elementos diagonais (variâncias) constantes e *a* representa os valores de covariância entre pares de safras ou idade.

Em espécies florestais, este modelo de análise é raramente aplicável. Isto porque, com o aumento dos caracteres de crescimento com a idade, as suposições de igual variância e correlação não são realistas. Neste caso, a padronização prévia dos dados remove apenas parcialmente a heterogeneidade de variâncias.

O modelo univariado de repetibilidade mais interação genótipos x medições é adequado em muitas situações e é discutido em detalhes no tópico 3.

c) **Modelo multivariado** (matriz de covariância não estruturada).

Este é o modelo mais completo, utiliza toda a informação simultaneamente e trata idades ou safras como sendo caracteres diferentes e correlacionados, considerando suas diferentes herdabilidades e correlações genéticas. A estrutura geral de correlação entre quatro diferentes idades é dada por:

$$R = \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix}$$, em que *a, b, c, d, e* e *f* representam diferentes valores de correlação.

Neste caso, a matriz R é denominada não estruturada, e o modelo é superparametrizado e proibitivo na prática quando muitas idades ou safras são consideradas. Sob modelo multivariado surgem problemas de estimação (matrizes não positivas definidas), principalmente quando o valor paramétrico encontra-se próximo ao limite do espaço do parâmetro.

d) **Modelo de Regressão Aleatória** (estrutura de covariância definida por funções de covariância).

Em um modelo de regressão aleatória (Meyer e Hill, 1997) os efeitos de tratamentos são modelados por $\sum_r^{l-1} \beta_{ir} \Phi(a_{ik}^*) r$, em que o termo β_{ir} denota o conjunto de l coeficientes de regressão aleatória para o i-ésimo tratamento, $\Phi(a_{ik}^*) r$ é o r-ésimo polinômio sobre a idade padronizada (a_{ik}^*) da medida k. A estimada matriz G para os efeitos de tratamentos é dada por G = ΦBΦ', em que Φ é a matriz contendo os efeitos aleatórios dos polinômios para as idades de medição e B é a estimada matriz de variância-covariância dos coeficientes de regressão polinomiais. Em termos genéricos, o valor fenotípico de um indivíduo é dado por $y_{ij} = F_{ij} + \sum_r^{l-1} \beta_{ir} \Phi(a_{ik}^*) r + \varepsilon_{ij}$, em que F_{ij} e ε_{ij} referem-se aos efeitos fixos do modelo e erros aleatórios, respectivamente.

O modelo de regressão aleatória pode ser considerado como um modelo multivariado reduzido e simplificado, o qual permite a obtenção dos mesmos parâmetros de interesse (herdabilidade em cada idade e correlação genética entre todos os pares de idade), que podem ser obtidos pelo modelo multivariado, porém com uma menor parametrização e com menor esforço computacional. Tal abordagem define diretamente funções de covariância contínuas e permite incluir na análise indivíduos com idades heterogêneas. A estrutura geral de covariância entre quatro diferentes idades é dada por:

$$G = \sum = \begin{bmatrix} z_{01} & z_{11} & z_{21} & z_{31} \\ z_{02} & z_{12} & z_{22} & z_{32} \\ z_{03} & z_{13} & z_{23} & z_{33} \\ z_{04} & z_{14} & z_{24} & z_{34} \end{bmatrix} \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix} \begin{bmatrix} z_{01} & z_{02} & z_{03} & z_{04} \\ z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ z_{31} & z_{32} & z_{33} & z_{34} \end{bmatrix}$$

$= \Phi B \Phi' = Q_i \Lambda_0 Q_i'$, em que:

z_{ij} = i-ésimo vetor polinomial ortogonal analisado na idade variável j.

B = Λ_0 = matriz de covariância dos regressores aleatórios ou coeficientes polinomiais.

$\Phi = Q_i$ = matriz com q+1 colunas contendo z_0, z_1, z_2,....., z_q, respectivamente, em que q é a ordem do polinômio e z_1 é o i-ésimo vetor polinomial ortogonal.

Um modelo de regressão aleatória pode ser ajustado somente com um intercepto e inclinação, com um termo quadrático adicional, com um termo cúbico adicional e assim sucessivamente de acordo com a ordem do polinômio de Legendre ajustado. Um exemplo simples (modelo somente com intercepto e inclinação) é apresentado por Resende (1999; 2002) e Resende et al. (2001b).

Os modelos de regressão aleatória modelam a trajetória dos valores genéticos no tempo, como desvios de outros efeitos fixos e aleatórios incluídos no modelo. Polinômios simples (de terceira ou quarta ordens) são utilizados para modelar esses desvios. Tais polinômios nem sempre são eficientes e o uso de funções mais flexíveis como polinômios de elevada ordem e splines cúbicas (sugeridas quando não se tem conhecimento prévio sobre o modelo biológico subjacente) podem conduzir a melhores estimativas.

A matriz de covariância genética G do modelo multivariado é reconstituída perfeitamente por G = ΦBΦ', desde que a ordem de B seja equivalente ao número de medidas repetidas. Este é o caso de um ajustamento completo, ou seja, é apenas uma parametrização diferente para o modelo multivariado. Entretanto, na prática, o interessante é obter uma redução na ordem da curva ajustada, de forma a simplificar o modelo multivariado e obter parcimônia. No modelo de regressão aleatória, B pode apresentar ordem baixa, dada por k = (p + 1), em que p é a ordem do polinômio ajustado. O valor k é a ordem da função de covariância e está associada à ordem do polinômio ajustado: k = 2 para o caso de uma regressão linear; k = 3 para o caso de uma função quadrática; k = 4 para o caso de uma função cúbica e assim sucessivamente. A matriz Φ apresenta ordem t x k em que t refere-se ao número de medições ou de idades.

e) **Modelo Auto-regressivo com Variâncias Heterogêneas** (Estrutura auto-regressiva).

Pletcher e Geyer (1999) sugeriram o uso dos modelos processo caráter para a análise de medidas repetidas. Esses modelos são baseados na teoria de processos estocásticos e foram estendidos por Jaffrezic e Pletcher (2000) visando relaxar a suposição mais restritiva de estacionariedade das correlações. O mais simples modelo processo caráter usa a função de covariância $C(t,s) = \sigma_t \sigma_s \rho^{(t-s)}$, em que $C(t,s)$ é a covariância entre medições nas safras ou tempos t e s, σ_t é o desvio padrão do caráter na safra t e $\rho^{(t-s)}$ é a correlação entre medições nas safras ou tempos t e s. Quando os dados são coletados em intervalos igualmente espaçados, esse processo caráter equivale ao modelo autoregressivo com variâncias heterogêneas (ARH). Assim, para medidas repetidas em plantas perenes, em que geralmente tem-se avaliações em intervalos regulares, os modelos de processo caráter e ARH são equivalentes. Adicionalmente, com variâncias constantes ou homogêneas através das medições e correlações exponenciais entre idades, os modelos de processo caráter e AR de ordem 1 (AR(1)) são equivalentes. A seguir é descrito o modelo ARH.

Este modelo assume correlações diferentes entre idades e reconhece a existência de correlação serial entre as medidas repetidas. A estrutura geral de correlação envolvendo quatro diferentes safras é definida como:

$$R = \begin{bmatrix} 1 & a^{|t_2-t_1|} & a^{|t_3-t_1|} & a^{|t_4-t_1|} \\ a^{|t_2-t_1|} & 1 & a^{|t_3-t_2|} & a^{|t_4-t_2|} \\ a^{|t_3-t_1|} & a^{|t_3-t_2|} & 1 & a^{|t_4-t_3|} \\ a^{|t_4-t_1|} & a^{|t_4-t_2|} & a^{|t_4-t_3|} & 1 \end{bmatrix}$$

Verifica-se que o modelo *AR* estima uma só correlação e projeta-a para os demais *lags*. Neste caso, a matriz de covariância genética é dada por *G* = *S* R *S* em que *S* é uma matriz diagonal com elementos equivalentes à raiz quadrada da variância genética em cada idade.

Em sua forma mais simples, o modelo processo caráter equivale ao tradicional modelo de repetibilidade (variâncias genéticas iguais e covariâncias fenotípicas iguais entre idades) e depende de apenas dois componentes de variância a serem estimados.

f) **Modelo com correlações específicas para cada intervalo de idade** (estrutura de covariância Toeplitz).

Este modelo considera correlações iguais para iguais intervalos entre idades e variâncias heterogêneas entre idades. É denominado modelo de correlação bandada e a matriz de correlação entre idades tem estrutura Toeplitz, que é dada por:

$$R = \begin{bmatrix} 1 & a & b & c \\ a & 1 & a & b \\ b & a & 1 & a \\ c & b & a & 1 \end{bmatrix}$$

Uma matriz Toeplitz apresenta várias diagonais (primárias, secundárias, terciárias etc.). A matriz de covariância genética é dada por *G* = *S* R *S,* em que R, no caso, é uma matriz Toeplitz.

As estruturas auto-regressivas (AR) e Toeplitz impõem fortes restrições às covariâncias, gerando também modelos mais parcimoniosos, assim como a regressão aleatória.

g) **Ajuste de uma curva *spline* cúbica no intervalo de idades considerado.**

Um polinômio ou função polinomial tem a forma

$Y = f(x) = a_0 + a_1 x + ...a_{n-1} x^{n-1} + a_n x^n$, em que:

x: variável do polinômio.

a_0 , ..., a_n : coeficientes do polinômio.

a_0 : coeficiente independente.

n: grau do polinômio ou expoente máximo da variável no polinômio.

Modelos polinomiais são úteis em situações em que o analista de dados sabe que os efeitos curvilineares estão presentes na verdadeira função resposta. São também úteis como funções aproximadoras para relações não lineares desconhecidas e complexas. Neste caso, o modelo polinomial é a expansão em série de Taylor da função desconhecida. No ajuste de modelos polinomiais, a ordem do modelo deve ser mantida tão baixa quanto possível.

Uma *spline* cúbica é uma curva alisada ou suavizada sobre um intervalo [a,b] formado pela ligação de segmentos polinomiais cúbicos (polinômios de ordem 3) a “pontos nó”, tais que a curva inteira e suas primeiras e segundas diferenças são contínuas sobre o intervalo. Refere-se a um processo de interpolação que propicia uma curva contínua e suave a partir dos pontos observados, sendo diferenciável e integrável para um domínio de interesse. Os nós são as abscissas dos pontos de junção dos segmentos. As *splines* cúbicas são desejáveis porque são de baixa ordem, levando a modelos mais parcimoniosos.

A função *spline* cúbica usa um processo polinomial de terceiro grau para a interpolação de valores entre cada par de pontos observados. Neste caso, uma polinomial diferente é usada para cada intervalo e cada uma é construída de forma a sempre passar pelos dados originais e apresentar uma derivada contínua nas junções entre cada intervalo. Uma curva definida por uma polinomial cúbica pode passar exatamente por quatro pontos, sendo que, no caso de uma longa seqüência de pontos, torna-se necessário usar uma sucessão de intervalos polinomiais. Para garantir que não ocorram mudanças súbitas nas inclinações ou curvaturas entre os intervalos sucessivos, tal função polinomial não é ajustada para quatro pontos e sim para dois.

A função *spline* cúbica foi desenvolvida inicialmente por Shoenberg (1946) e aperfeiçoada por Walsh et al. (1962). Verbyla et al. (1999) e White et al. (1999) sugeriram o uso de funções mais flexíveis, tais quais polinomiais de elevada ordem ou *splines* cúbicas para modelar dados longitudinais quando não existe conhecimento prévio sobre o modelo biológico adjacente ao caráter.

Splines cúbicas podem ser incorporadas ao contexto dos modelos lineares mistos (White et al., 1999; Verbyla et al., 1999). Nesse caso, os efeitos de tratamento são modelados por $b_{i0} + b_{i1}t_{ik} + \sum_{l=2}^{q-1} b_{il} z_l(t_{ik})$, em que b_{i0} denota o intercepto para o tratamento i, b_{i1} denota a inclinação para o tratamento i e b_{il} denota o coeficiente de regressão aleatória para o i-ésimo tratamento no nó l. O termo t_{ik} denota a idade da avaliação e $z_l(t_{ik})$ representa o coeficiente da *spline* para a idade t_{ik}. A estimação da matriz G para os efeitos de tratamento é dada por G = ΩZΩ', em que Ω é uma matriz contendo os efeitos aleatórios da spline para as idades de avaliação e Z é a matriz estimada de variância-covariância dos coeficientes da *spline*. Exemplos e detalhes da obtenção de *splines* cúbicas são apresentados por Davis (1986) e Green e Silverman (1994). Outra alternativa é o uso da *spline* B, conforme Meyer (2005) e Welham et al. (2005; 2006).

h) **Modelo Ante Dependência Estruturado (SAD)**

A idéia básica dos modelos ante-dependência (AD) é que uma observação no tempo t pode ser explicada por observações prévias. Os valores da observação em determinado tempo depende da observação no tempo imediatamente anterior e mais uma nova variação, ou seja, mais um acréscimo ocorrido no intervalo entre as duas medições. Nunez-Anton e Zimmerman (2000) propuseram modelo ante-dependência estruturado (SAD) no qual o número de parâmetros é menor do que nos tradicionais modelos AD. Esses modelos podem lidar com padrões de correlação altamente não estacionários e correspondem, em suas especificações mais simples, a uma generalização dos modelos autoregressivos. Consideram também a heterogeneidade de variância entre medições. O SAD é também favorável em termos do número de parâmetros ajustados o qual é geralmente pouco maior do que o número ajustado pelos modelos ARH. A matriz de covariância é da forma:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_4\rho_2\rho_3 \\ Sim. & & \sigma_3^2 & \sigma_3\sigma_4\rho_3 \\ & & & \sigma_4^2 \end{bmatrix}.$$

As estruturas CS e CSH são discutidas nos tópicos 3 e 4.

2 COMPARAÇÃO ENTRE ESTRUTURAS DE COVARIÂNCIA NA ANÁLISE DE MEDIDAS REPETIDAS

Estas alternativas (multivariado, ARH, SAD, regressão aleatória) de modelagem podem ser aplicadas aos vários fatores aleatórios do modelo estatístico base. No contexto da estatística experimental com efeitos de tratamentos considerados fixos, estas modelagens em geral se aplicam somente aos resíduos. Mas no caso do melhoramento, em que os tratamentos genéticos são considerados de efeitos aleatórios, essas modelagens podem ser aplicadas aos efeitos residuais e também genéticos. Inclusive, diferentes modelagens podem ser empregadas para os efeitos genéticos e residuais. Para algumas espécies, a modelagem dos efeitos genéticos como ARH ou SAD e dos erros por um modelo multivariado, mostraram-se eficientes (Resende, Thompson e Welham, 2003). A seguir, são apresentadas comparações entre algumas dessas modelagens, com base em Resende e Thompson (2003) e Resende et al. (2006).

Os resultados seguintes referem-se à avaliação de três colheitas anuais de massa verde em 1.800 indivíduos de erva-mate, avaliados em um experimento em blocos ao acaso com cinco repetições e seis plantas por parcela, perfazendo um total de 5.400 observações. Na Tabela 35 são apresentados resultados referentes ao modelo multivariado completo (matriz de covariância não estruturada) aplicado sobre os dados originais. Os efeitos de blocos, colheitas e interação blocos x colheitas foram ajustados como fixos.

Tabela 35. Estimativas dos parâmetros de variância e covariância associados ao modelo multivariado completo aplicado aos dados originais de três colheitas. Número total de parâmetros igual a 18.

Tratamento (genético)			**Parcela**			**Resíduo**		
Covar.\Variância\Correl.			**Covar.\Variância\Correl.**			**Covar.\Variância\Correl.**		
0.0134	0.9239	0.9984	0.0248	0.8638	0.7095	0.1011	0.6766	0.6128
0.0342	0.1020	0.9211	0.0422	0.0964	0.9123	0.1495	0.4827	0.7686
0.0673	0.1711	0.3380	0.0817	0.2072	0.5352	0.2755	0.7551	1.9990
Deviance = -787.172								

O valor de *deviance* (Tabela 35) revela que o modelo multivariado completo é mais adequado do que o modelo de repetibilidade aplicado sobre os dados originais (*deviance* de 5070.64, resultados não mostrados). O modelo multivariado apresenta altos valores de correlação genética entre pares de colheitas. As correlações encontram-se todas dentro do espaço paramétrico, mas o modelo teve que ser restrito para se conseguir isto. Sem restringir a matriz G para ser positiva definida, foram obtidas correlações maiores que 1 e estimativas negativas de componentes de variância. No modelo restrito, a matriz G é distorcida e esse processo envolve regressar as variâncias em direção às suas médias. A análise não restrita é menos viciada porque vício é introduzido quando se restringe as soluções ao espaço paramétrico. A convergência foi mais difícil à medida que o número de parâmetros aumentou. Então, modelos mais adequados precisaram ser pesquisados.

Resultados referentes ao modelo processo caráter, denominado autoregressivo de primeira ordem com variâncias heterogêneas (ARH), para os efeitos de tratamentos, são apresentados na Tabela 36. Os demais fatores, parcela e resíduo, foram ajustados pelo modelo multivariado completo.

Tabela 36. Estimativas dos parâmetros de variância e covariância associados ao modelo processo caráter, denominado autoregressivo de primeira ordem com variâncias heterogêneas (ARH), aplicado aos dados originais de três colheitas. Número total de parâmetros igual a 16.

Tratamento (genético)			**Parcela**			**Resíduo**		
Covar.\Variância\Correl.			**Covar.\Variância\Correl.**			**Covar.\Variância\Correl.**		
0.0129	0.9761	0.9528	0.0254	0.8667	0.7532	0.1011	0.6766	0.6128
0.0357	0.1033	0.9761	0.0430	0.0968	0.9109	0.1495	0.4827	0.7686
0.0619	0.1792	0.3261	0.0891	0.2104	0.5510	0.2755	0.7551	1.9990
Deviance = -782.94								

Os modelos ARH e multivariado apresentaram quase a mesma *deviance* e os valores de AIC (critério de informação de Akaike) foram -750.94 e -751.17, respectivamente, os quais são basicamente o mesmo -751. Então, os dois modelos são equivalentes pelo critério de parcimônia. Entretanto, o ARH apresentou fácil convergência sem a necessidade de restringir a matriz G para ser positiva definida. Adicionalmente, ajustou um menor (dois a menos) número de parâmetros que o modelo multivariado, retornou correlações dentro do espaço paramétrico e propiciou uma correlação mais realista entre as duas colheitas mais distantes (1 e 3). O modelo ARH é então muito superior e assume estacionariedade e mesmas correlações em intervalos de colheitas de mesmo tamanho.

Outro modelo avaliado foi o de antedependência estruturado (SAD), o qual também apresenta parcimônia e não assume estacionariedade. Os resultados são apresentados na Tabela 37. Os demais fatores, parcela e resíduo, foram ajustados pelo modelo multivariado completo.

Tabela 37. Estimativas dos parâmetros de variância e covariância associados ao modelo ante dependência estruturado (SAD), aplicado aos dados originais de três colheitas. Número total de parâmetros igual a 17.

Tratamento (genético)			**Parcela**			**Resíduo**		
Covar.\Variância\Correl.			**Covar.\Variância\Correl.**			**Covar.\Variância\Correl.**		
0.0128	0.9840	0.9580	0.0254	0.8667	0.7532	0.1011	0.6766	0.6128
0.0358	0.1032	0.9730	0.0430	0.0968	0.9109	0.1495	0.4827	0.7686
0.0618	0.1784	0.3250	0.0891	0.2104	0.5510	0.2755	0.7551	1.9990
Deviance = -783.04								

Os modelos SAD e ARH apresentaram basicamente a mesma *deviance* (-783) e então são equivalentes por esse critério. No entanto, o modelo SAD ajustou um parâmetro a mais que o ARH e por isso não é preferido em termos de parcimônia pelo critério de informação de Akaíke (AIC). Os resultados associados aos efeitos de parcela e residuais foram exatamente iguais pelos dois modelos. Os parâmetros associados ao componente genético foram ligeiramente diferentes mas ambos são coerentes em termos de magnitude dos coeficientes de correlação, isto é, valores menores para o intervalo 1-3. Isto não foi observado no modelo multivariado completo. Ambos os modelos poderiam ser usados eficientemente na prática. O modelo SAD permite diferentes correlações para intervalos de mesmo tamanho.

Essas duas classes de modelo foram também usados para modelagem de outros termos aleatórios do modelo. Os resultados referentes às correlações associadas aos fatores tratamento e parcela, modelados por ARH e SAD, são apresentados na Tabela 38. Os efeitos residuais foram ajustados pelo modelo multivariado completo.

Tabela 38. Estimativas das correlações associadas aos modelos ante dependência estruturado (SAD) e processo caráter (ARH) para a modelagem de ambos os fatores, tratamento e parcela, aplicados aos dados originais de três colheitas. Número total de parâmetros igual a 16 para o SAD e 14 para o ARH.

Tratamento (genético)			**Parcela**			**Resíduo**		
ARH\SAD			**ARH\SAD**			**ARH\SAD**		
-	0.990	0.968	-	0.851	0.769	-	0.6766	0.6128
0.982	-	0.977	0.882	-	0.903	0.6766	-	0.7686
0.964	0.982	-	0.778	0.882	-	0.6128	0.7686	-
Deviance ARH\SAD = -780.76\-782.20								

Os resultados mostram que os efeitos de parcela podem ser perfeitamente modelados pelos processos ARH ou SAD. Os valores de deviance foram próximos aos prévios, obtidos quando os efeitos de parcela foram modelados via o processo multivariado completo. Os valores de AIC aqui foram de –752.76 e –750.20 para ARH e SAD, respectivamente, os quais são próximos aos valores –751 e –749 para ARH e SAD, respectivamente, obtidos com os dois modelos mas com efeito de parcela multivariado. Comparando esses quatro valores de AIC, a escolha é pelo modelo ARH para ambos os fatores, tratamento e parcela (AIC = –752.76).

A modelagem do termo residual via ARH foi também avaliada. A *deviance* resultante para a modelagem dos três fatores simultaneamente por processos ARH foi de –677.46, a qual é alta em comparação com a modelagem anterior. Também as correlações residuais obtidas foram muito diferentes das relatadas anteriormente. Então, os efeitos residuais devem ser modelados pelo processo multivariado completo.

Outras abordagens foram também avaliadas. O modelo associado à estrutura de correlação bandada ou Toeplitz convergiu com uma *deviance* de –794.28. No entanto, forneceu uma correlação genética maior que 1, justamente a correlação (entre as colheitas 1 e 3) que deveria ser a menor delas. Esse modelo assume iguais correlações para intervalos de mesmo tamanho tal como o modelo ARH, mas os elementos das várias diagonais são diferentes e não são função da correlação para o intervalo 1.

Modelos de regressão aleatória para os efeitos de tratamentos foram também avaliados e os resultados são apresentados na Tabela 39 para os modelos completo e reduzido. Os demais fatores, parcela e resíduo, foram ajustados pelo modelo multivariado completo.

Tabela 39. Estimativas dos parâmetros de variância e covariância associados ao modelo de regressão aleatória aplicado aos dados originais de três colheitas.

Modelo completo (ajuste quadrático) com número total de parâmetros igual a 18

Tratamento (genético)			**Parcela**			**Resíduo**		
Covar.\Variância\Correl.			**Covar.\ Variância\Correl.**			**Covar.\Variância\Correl.**		
0.0134	0.9239	0.9984	0.0248	0.8638	0.7095	0.1011	0.6766	0.6128
0.0342	0.1020	0.9211	0.0422	0.0964	0.9123	0.1495	0.4827	0.7686
0.0673	0.1711	0.3380	0.0817	0.2072	0.5352	0.2755	0.7551	1.9990
Deviance = -787.172								

Modelo reduzido (ajuste linear) com número total de parâmetros igual a 15

Tratamento (genético)			Parcela			Resíduo		
Covar.\Variância\Correl.			Covar.\Variância\Correl.			Covar.\Variância\Correl.		
0.0098	1.0552	1.0765	0.0248	0.8638	0.7095	0.1011	0.6766	0.6128
0.0331	0.1004	1.0040	0.0422	0.0964	0.9123	0.1495	0.4827	0.7686
0.0563	0.1677	0.2791	0.0817	0.2072	0.5352	0.2755	0.7551	1.9990
Deviance = -777.08								

Os resultados foram idênticos àqueles (os quais também não foram adequados) do modelo multivariado, conforme esperado para o modelo completo de regressão aleatória, isto é, para o modelo em que se ajustou um polinômio quadrático. Em busca de parcimônia, um modelo reduzido foi avaliado. A *deviance* (-777) desse modelo é maior do que aquela (-783) obtida com os modelos ARH e SAD para os efeitos de tratamentos (Tabelas 38 e 39). O valor de AIC é –747, o qual é maior do aqueles obtidos com os processos ARH (-751) e SAD (-749). Então, o modelo reduzido de regressão aleatória não é o escolhido. Também, esse modelo mostrou uma pobre reconstrução da matriz G para os efeitos de tratamentos, conduzindo todas as correlações a valores maiores que 1

(Tabela 39). Esses resultados concordam com Apiolaza et al. (2000) que concluíram que os modelos de regressão aleatória foram inapropriados para descrever a estrutura de covariância envolvendo o crescimento de *Pinus radiata* ao longo das idades.

O ajuste de *splines* cúbicas foi também avaliado. A *deviance* obtida foi alta (–748.33), a qual foi a pior dentre todos os modelos parcimoniosos avaliados. Esse resultado era esperado devido ao pequeno número de idades disponíveis para o ajuste. Em resumo, as melhores modelagens foram a ARH e SAD para ambos os fatores, tratamento e parcela. Esses modelos podem ser estendidos para incorporar estruturas espaciais para os resíduos.

3 ANÁLISE DE MEDIDAS REPETIDAS EM EXPERIMENTOS INDIVIDUAIS

A análise de experimentos de medidas (ou colheitas) repetidas nas mesmas parcelas e/ou indivíduos apresenta peculiaridades pelo fato das várias colheitas serem correlacionadas entre si e pela possibilidade de haver heterogeneidade de variâncias e de covariâncias entre as várias colheitas ou safras. Um modelo completo e ótimo para analisar um conjunto de dados dessa natureza é o modelo multivariado (também denominado modelo com matriz de covariância não estruturada entre colheitas, UN) o qual trata cada colheita como se fosse uma variável diferente. E essa estrutura de covariância é aplicada a todos os fatores aleatórios do modelo tais quais os efeitos genotípicos de tratamentos, efeitos de parcela e os efeitos residuais. Porém, considerando um número relativamente grande (três ou mais) de colheitas, tal modelo é difícil de ser ajustado (apresentando problema de convergência), além de ser superparametrizado, ou seja, depender da estimativa de um grande número de parâmetros.

Como apresentado no Capítulo 8, a modelagem mais simples para cada fator é via estrutura de covariância de simetria composta (CS), a qual assume tanto homogeneidade de variâncias quanto de covariâncias entre colheitas. Essa abordagem é desejável porque depende do menor número possível de estimativas de parâmetros. No entanto, pode ser ineficiente no caso de grande heterogeneidade de variância e covariância entre colheitas, algumas vezes decorrentes de efeitos de escala das observações de uma colheita a outra. Adotar a modelagem via estrutura CS para os

fatores genótipos, parcela e resíduos é exatamente igual à abordagem de um modelo univariado com os fatores genótipos, blocos e colheitas (ou medições) e suas interações duplas e tripla. Para uma aplicação segura dessa estrutura CS, recomenda-se, em caso de presença de grande heterogeneidade de variâncias, a correção prévia dos dados de cada colheita por meio da multiplicação dos mesmos pela razão hi/hm, em que hi refere-se à raiz quadrada herdabilidade na colheita i e hm refere-se à raiz quadrada da média das herdabilidades nas várias colheitas, conforme Resende (2004). Esse procedimento considera simultaneamente a heterogeneidade de variância genética e fenotípica entre colheitas, penalizando pela alta variância fenotípica e capitalizando pela alta variância genética, o que equivale aproximadamente a padronizar pelo desvio padrão ambiental. A padronização pode ser feita também diretamente dividindo os dados pelo desvio padrão ambiental de cada colheita e multiplicando pela média desses desvios padrões. Procedendo-se dessa forma, a aplicação da estrutura CS fornece resultados semelhantes aqueles que se obtém quando se aplica a estrutura de simetria composta com variâncias heterogêneas (CSH).

No caso de análises conjuntas de colheitas duas a duas e com presença de homogeneidade de variâncias (ou corrigindo a heterogeneidade), a aplicação das estruturas UN e CS produzem resultados semelhantes. Ainda no caso de análises de colheitas duas a duas, com heterogeneidade de variâncias, as estruturas UN e CSH são idênticas. Com análises envolvendo mais que duas colheitas e com presença de homogeneidade de variâncias (ou corrigindo a heterogeneidade), as estruturas UN e CS produzem resultados semelhantes. Em resumo, com correção para a heterogeneidade de variâncias e em presença de correlações genéticas não muito discrepantes entre pares de colheitas não há necessidade de se usar o modelo multivariado ou estrutura UN.

Com correlações genéticas muito discrepantes entre pares de colheitas e grande número de colheitas, outras estruturas intermediárias entre UN e CS podem ser usadas. Dentre essas, citam-se: fator analítica multiplicativa sob modelos mistos (FAMM) (Resende e Thompson, 2003; 2004); componentes principais genéticos sob modelos mistos (PCAM) (Meyer e Kirkpatrick, 2005); autoregressiva com variâncias heterogêneas (ARH); antedependência estruturada (SAD); Toeplitz ou de correlação bandada; modelos de regressão aleatória (RR); ajuste de *splines*. Dentre os modelos citados, os mais adequados a estudos de medidas repetidas no tempo são: ARH, SAD, Toeplitz, RR, PCAM e *Splines* (Resende, 2002, p.522; Resende, 2004). No entanto, os modelos

ajustados via RR, PCAM e *Splines* tendem a ser pouco efetivos no melhoramento de plantas devido ao pequeno número de colheitas praticado em cada genótipo, conforme demonstrado anteriormente. Tais técnicas são muito empregadas no melhoramento animal, em que indivíduos de diferentes idades são avaliados produzindo um grande número de pontos em termos de idades. Essas técnicas são mais adequadas para avaliar caracteres infinitamente dimensionais propriamente ditos.

Dessa forma, para caracteres de plantas perenes, em que as correlações seguem alguns padrões em função das distâncias entre as medidas consideradas, em geral havendo decréscimo das correlações em função do aumento dessas distâncias, os modelos ARH e SAD tendem a ser os mais adequados (Resende e Thompson, 2003; Resende et al., 2006). Assim como o modelo ARH, o Toeplitz pode também ser adequado nas situações em que as correlações apresentam a mesma magnitude em intervalos de idade de mesma dimensão, por exemplo, correlações iguais entre as colheitas 1 e 2 e entre as colheitas 2 e 3. Entretanto, o modelo Toeplitz é mais parametrizado do que o ARH. Mas em geral, com a estabilização do caráter, a correlação entre as idades 2 e 3 tende a ser maior que a correlação entre as colheitas 1 e 2, e portanto, o modelo SAD parece ser mais coerente pois contempla isso.

Em resumo, com correção para heterogeneidade de variâncias, a estrutura CS pode ser aplicada com sucesso, a menos que haja uma grande heterogeneidade das correlações entre as colheitas. Nesse caso, as estruturas ARH e SAD podem ser aplicadas com vantagens, embora esses modelos sejam mais parametrizados (critérios como o AIC e BIC podem ser usados para verificar se essa maior parametrização é compensadora). Modelos muito parametrizados apresentam um alto custo em termos de consumo de graus de liberdade. Isso conduz a menor número de graus de liberdade para estimar cada componente de variância e, portanto, conduzem a estimativas menos precisas, fato que pode comprometer as predições baseadas nesses modelos. No entanto, se o caráter é o mesmo de uma colheita para outra, espera-se correlações mais homogêneas entre colheitas e, portanto, grande utilidade do modelo CS.

O modelo CS ou modelo completo de repetibilidade associado ao delineamento experimental de blocos ao acaso com uma observação por parcela é dado por $Y_{ijk} = \mu + g_i + b_j + m_k + gb_{ij} + gm_{ik} + bm_{jk} + gbm_{ijk}$, em que u é o efeito da média geral, gi é o efeito do genótipo i, b_j é o efeito do bloco j, m_k é o efeito da medição ou colheita k, gm_{ik} é o efeito da

interação genótipos x medições, bm_{jk} é o efeito da interação blocos x medições, gb_{ij} é o efeito da interação genótipos x blocos e gbm_{ijk} é a interação tripla ou resíduo aleatório. Os efeitos gb_{ij} referem-se aos efeitos de parcela, os quais são efeitos de ambiente permanente de uma colheita para outra.

Esse modelo difere consideravelmente do modelo apresentado no Capítulo 8 para a análise da interação genótipo x ambiente, dado por $Y_{ijk} = u + g_i + b_{j/k} + l_k + gl_{ik} + e_{ijk}$, em que u é o efeito da média geral, gi é o efeito do genótipo i, $b_{j/k}$ é o efeito do bloco j dentro do local k, l_k é o efeito do local k, gl_{ik} é o efeito da interação genótipos x locais e e_{ijk} é o erro ou resíduo aleatório. Embora a colheita possa ser tomada como ambiente, não há casualização dos blocos e genótipos nas diferentes colheitas, ou seja, não há casualização das parcelas, isto porque os blocos não são hierárquicos em relação às colheitas, como o são em relação aos locais. Na presente situação, blocos e colheitas apresentam classificação cruzada e portanto existe o efeito bm_{jk}. Também, os efeitos permanentes de parcela (gb_{ij}) não são contemplados no modelo da análise conjunta de locais. Portanto, os modelos são completamente diferentes. Este tem sido um erro comum em vários artigos com plantas perenes, em que as colheitas são assumidas como ambientes e o modelo com blocos hierárquicos a ambientes utilizado erroneamente. A inclusão dos efeitos permanentes de parcelas é essencial para considerar e eliminar os efeitos da correlação residual entre medidas repetidas.

Considerando os efeitos ambientais de blocos (b), medições (m) e interação blocos x medições como fixos (pois são efeitos ambientais para os quais os dados devem ser corrigidos) no modelo completo de repetibilidade ou CS apresentado acima, os mesmos podem ser ajustados somados a média geral, em um único vetor de efeitos fixos (β) dado pela combinação bloco-medição. Assim, o modelo linear misto resultante equivale a $Y_{ijk} = \beta_{jk} + g_i + gm_{ik} + gb_{ij} + gbm_{ijk}$. Esse modelo foi implementado nos modelos 55 (blocos ao acaso) e 78 (látice) do Selegen e são ótimos quando os dados são corrigidos previamente quanto à heterogeneidade de variâncias e quando as correlações entre pares de colheitas não são muito discrepantes.

Desdobrando este modelo em termos de efeitos permanentes (p) e temporários (t), tem-se $y = \beta + g_p + g_t + p_p + p_t$, em que:

$g_i = g_p$: efeito de genótipo, permanente através das colheitas.

$gm_{ik} = g_t$: efeito de genótipo, temporário em cada colheita.

$gb_{ij} = p_p$: efeito de parcela, permanente através das colheitas.

$gbm_{ijk} = p_t$: efeito de parcela, temporário em cada colheita.

Em termos de variâncias destes efeitos, tem-se:

$\sigma^2_{gp} = \sigma^2_g$: variância genotípica ou covariância dos efeitos genotípicos através das colheitas; é a covariância genotípica entre colheitas em um modelo multivariado (estrutura UN).

$\sigma^2_{gt} = \sigma^2_{gm}$: variância da interação genótipos x medições.

$\sigma^2_{pp} = \sigma^2_{gb}$: variância dos efeitos permanentes de parcela ou covariância dos efeitos de parcela através das colheitas em um modelo multivariado.

$\sigma^2_{pt} = \sigma^2_{gbm}$: variância dos efeitos temporários de parcela ou da interação parcelas x medições.

Verifica-se que tal modelo é bastante próximo ao modelo multivariado (estrutura UN para os efeitos genotípicos, contemplando simultaneamente os efeitos g_i e gm_{ik} e estrutura UN para os efeitos de parcela ou residuais contemplando simultaneamente os efeitos gb_{ij} e gbm_{ijk}), desde que haja homogeneidade de variâncias. Esse modelo completo de repetibilidade é também denominado modelo de repetibilidade + interação genótipos x medições ou modelo com estrutura CS. Assumindo que a correlação genotípica através das medições aproxima 1 (ou seja, assumindo que o caráter é o mesmo de uma colheita para outra), o modelo se reduz a $y = \beta + g_p + p_p + p_t = \beta + g + gb + gbm$, o qual é denominado modelo simplificado de repetibilidade ou modelo de repetibilidade.

Com avaliação em nível de plantas individuais e várias plantas por parcela, o modelo deve ser estendido. O modelo completo de repetibilidade associado ao delineamento experimental de blocos ao acaso com várias plantas por parcela é dado por $Y_{ijkl} = \mu + g_i + b_j + m_k + gb_{ij} + gm_{ik} + bm_{jk} + gbm_{ijk} + e_{ijkl}$. Considerando os efeitos ambientais de blocos (b), medições (m) e interação blocos x medições como fixos, os mesmos podem ser

ajustados somados a média geral, em um único vetor de efeitos fixos (β) dado pela combinação bloco-medição. Assim, o modelo linear misto resultante equivale a $Y_{ijkl} = \beta_{jk} + g_i + gm_{ik} + gb_{ij} + gbm_{ijk} + e_{ijkl}$.

Desdobrando este modelo em termos de efeitos permanentes (p) e temporários (t), tem-se $y = \beta + g_p + g_t + p_p + p_t + e_p + e_t$, em que:

$g_i = g_p$: efeito de genótipo, permanente através das safras.

$gm_{ik} = g_t$: efeito de genótipo, temporário em cada safra.

$gb_{ij} = p_p$: efeito de parcela, permanente através das safras.

$gbm_{ijk} = p_t$: efeito de parcela, temporário em cada safra.

$e_{ijk} = e_p + e_t$: efeito permanente + temporário de indivíduo dentro de parcela.

Em termos de variâncias destes efeitos tem-se:

$\sigma^2_{gp} = \sigma^2_g$: variância genotípica ou covariância dos efeitos genotípicos através das safras; é a covariância genotípica através das safras em um modelo multivariado.

$\sigma^2_{gt} = \sigma^2_{gm}$: variância da interação genótipos x medições.

$\sigma^2_{pp} = \sigma^2_{gb}$: variância dos efeitos permanentes de parcela ou covariância dos efeitos de parcela através das safras em um modelo multivariado.

$\sigma^2_{pt} = \sigma^2_{gbm}$: variância dos efeitos temporários de parcela ou da interação parcelas x medições.

σ^2_{ep}: variância permanente de indivíduo dentro de parcela ou covariância dos efeitos de indivíduos dentro de parcela através das safras em um modelo multivariado.

σ^2_{et}: variância temporária de indivíduo dentro de parcela.

Verifica-se que tal modelo é bastante próximo ao modelo multivariado, desde que haja homogeneidade de variâncias. Assumindo que a interação dos ambientes das parcelas x medições é desprezível e/ou pode ser reunido ao erro temporário, o modelo simplifica-se para $y = \beta + g_p + g_t + p_p + e_p + e_t = \beta + g + gm + gb + e_p + e_t$, o qual é denominado modelo individual de repetibilidade + interação genótipos x medições. Assumindo adicionalmente que a correlação genotípica através das medições aproxima 1, o modelo se reduz a $y = \beta + g_p + p_p + e_p + e_t = \beta + g + gb + e_p + e_t$, o qual é denominado modelo individual simplificado de repetibilidade. O modelo g + ge para análise de múltiplos experimentos também deriva do modelo completo de repetibilidade, mudando-se a dimensão tempo (t) para a dimensão espaço (e). Assim, $y = \beta + g_p + g_e + p_p + p_e + e_p + e_e$. Como não existe covariância dos efeitos ambientais através do espaço, $p_p = e_p = 0$, e o modelo simplifica-se para $y = \beta + g + g_e + p_e + e_e$, em que $g_e, p_e \, e \, e_e$ significam efeitos específicos para cada espaço ou ambiente.

Para a estimação da repetibilidade em experimentos com delineamento experimental, modelos que estimem simultaneamente a herdabilidade, a repetibilidade e a correlação genética através das colheitas, devem ser empregados. Modelos simples que estimam a repetibilidade, ignorando a estrutura experimental, os efeitos de blocos, os efeitos de parcela e o parentesco genético entre os indivíduos em avaliação, não devem ser usados.

4 ANÁLISE DE MEDIDAS REPETIDAS EM MÚLTIPLOS EXPERIMENTOS

A avaliação de genótipos em vários locais e em várias colheitas ou safras em plantas perenes gera dados com simultânea dependência através dos locais e do tempo. Estruturas de covariância para modelar este tipo de base de dados não são facilmente encontradas. Isto porque, em geral, nesse caso as correlações entre medidas repetidas dentro de locais são de alta magnitude, mas as correlações entre medidas através dos locais podem ser de baixa magnitude, fato que dificulta a modelagem do fator tratamentos. Assim sendo, as estruturas ARH, SAD e Toeplitz mencionadas nos tópicos anteriores não são adequadas. A estrutura FAMM mostrou-se

adequada nesse caso, embora tenha havido dificuldade de convergência no processo iterativo envolvido na análise, em alguns casos (Resende e Thompson, 2003; 2004). Nesse caso, Gilmour (2006) sugere o uso das estruturas FAMM, AR e Cholesky com reduzido número de parâmetros. No entanto, relata a dificuldade de se considerar efetivamente medidas repetidas em vários caracteres simultaneamente, em uma única matriz de covariância.

Assim, estruturas do tipo simetria composta (CS) com correção para heterogeneidade de variâncias (propiciando resultados semelhantes aos obtidos pela estrutura simetria composta com variâncias heterogêneas - CSH) podem ser aplicadas de maneira satisfatória.

O modelo CS ou modelo simultâneo de repetibilidade, herdabilidade, interação genótipos x locais, interação genótipos x colheitas e interação tripla, associado ao delineamento experimental de blocos ao acaso com uma observação por parcela é dado por :

$$Y_{ijkn} = \mu + g_i + b_{j/n} + m_k + l_n + gb_{ij/n} + gm_{ik} + gl_{in} + bm_{jk/n} + ml_{kn} + gml_{ikn} + gbml_{ijk/n},$$

em que u é o efeito da média geral, gi é o efeito do genótipo i, $b_{j/n}$ é o efeito do bloco j dentro do local n, m_k é o efeito da medição ou colheita k, l_n é o efeito do local n, gm_{ik} é o efeito da interação genótipos x medições, gl_{in} é o efeito da interação genótipos x locais, $bm_{jk/n}$ é o efeito da interação blocos x medições dentro de locais, $gb_{ij/n}$ é o efeito da interação genótipos x blocos dentro de locais, ml_{kn} é o efeito da interação medições x locais, gml_{ijn} é a interação tripla genótipos x medições x locais, $gbml_{ijk/n}$ é o resíduo aleatório. Os efeitos gb_{ij} referem-se aos efeitos de parcela/locais, os quais são efeitos de ambiente permanente de uma colheita para outra. Os efeitos gl_{in} também são efeitos de ambiente permanente de uma colheita para outra.

Esse modelo difere consideravelmente do modelo (implementado como modelo 114 no Selegen-Reml/Blup) utilizado em culturas anuais, para a análise envolvendo a avaliação de genótipos em vários locais e anos de plantio, dado por $Y_{ijkn} = \mu + g_i + b_{j/k/n} + a_k + l_n + ga_{ik} + gl_{in} + al_{kn} + gal_{ikn} + gbal_{ij/k/n}$, em que u é o efeito da média geral, gi é o efeito do genótipo i, $b_{j/k/n}$ é o efeito do bloco j dentro do ano k dentro do local n, l_n é o efeito do local n, a_k é o efeito do ano de plantio k, gl_{in} é o efeito da interação genótipos x locais, ga_{ik} é o efeito da interação genótipos x anos de plantio, al_{kn} é o efeito da interação locais x anos de plantio, gal_{ikn} é o efeito da interação genótipos x anos x locais, e $gbal_{ij/k/n}$ é o erro ou resíduo aleatório.

Embora a colheita possa ser tomada como anos, não há casualização dos blocos e genótipos nas diferentes colheitas, ou seja, não há casualização das parcelas. Isto porque os blocos não são hierárquicos em relação às colheitas, como o são em relação aos anos de plantio. Na presente situação, blocos e colheitas apresentam classificação cruzada dentro de locais e portanto existe o efeito $bm_{jk/n}$. Também, os efeitos permanentes de parcela dentro de locais ($gb_{ij/n}$) não são contemplados no modelo da análise conjunta de locais e anos de plantio. Portanto, os modelos são completamente diferentes. Este tem sido um erro comum em vários artigos com plantas perenes, em que as colheitas são assumidas como anos de plantio e o modelo para análise envolvendo a avaliação de genótipos em vários locais e anos de plantio (com blocos hierárquicos a anos de plantio) utilizado erroneamente. A inclusão dos efeitos permanentes de parcelas dentro de locais é essencial para considerar e eliminar os efeitos da correlação residual entre medidas repetidas.

Considerando os efeitos ambientais de blocos dentro de locais (b), medições (m), locais (l) e as interações blocos x medições dentro de locais e locais x medições como efeitos fixos (pois são efeitos ambientais para os quais os dados devem ser corrigidos) no modelo completo de repetibilidade ou CS apresentado acima, os mesmos podem ser ajustados somados a média geral, em um único vetor de efeitos fixos (β) dado pela combinação bloco-medição-local. Assim, o modelo linear misto resultante equivale a $Y_{ijkn} = \beta_{jkn} + g_i + gm_{ik} + gl_{in} + gb_{ij/n} + gml_{ijn} + gbml_{ijk/n}$. Esse modelo CS foi implementado no modelo 155 do *software* Selegen-Reml/Blup. Este elegante modelo de análise propicia a estimação simultânea da herdabilidade, da repetibilidade, das correlações genéticas através das colheitas, dos locais, das colheitas e locais, das colheitas para um dado local, dos locais para uma dada colheita, das colheitas para a média de locais e da correlação genética através dos locais para a média de colheitas, mesmo sob desbalanceamento.

Desdobrando este modelo em termos de efeitos permanentes (p) e temporários (t), tem-se $y = \beta + g_p + g_t + g_{pn} + p_p + g_{tn} + p_t$, em que:

$g_i = g_p$: efeito de genótipo, permanente através das colheitas e locais.

$gm_{ik} = g_t$: efeito de genótipo, temporário em cada colheita.

$gl_{in} = g_{pn}$: efeito de genótipo, permanente através das colheitas em cada local mas não permanente através dos locais.

$gml_{ikn} = g_{tn}$: efeito de genótipo, temporário em cada colheita e não permanente através dos locais, ou seja, peculiar a cada safra e local.

$gb_{ij/n} = p_p$: efeito de parcela dentro de locais, permanente através das colheitas.

$gbml_{ijk/n} = p_t$: efeito de parcela dentro de locais, temporário em cada colheita.

Em termos de variâncias destes efeitos, tem-se:

$\sigma^2_{gp} = \sigma^2_g$: variância genotípica ou covariância dos efeitos genotípicos através das colheitas e locais; é a covariância genotípica através das colheitas e locais em um modelo multivariado (estrutura UN).

$\sigma^2_{gt} = \sigma^2_{gm}$: variância da interação genótipos x medições.

$\sigma^2_{gpn} = \sigma^2_{gl}$: variância da interação genótipos x locais.

$\sigma^2_{gtn} = \sigma^2_{gml}$: variância da interação genótipos x medições x locais.

$\sigma^2_{pp} = \sigma^2_{gb}$: variância dos efeitos permanentes de parcela dentro de locais ou covariância dos efeitos de parcela dentro de locais através das colheitas.

$\sigma^2_{pt} = \sigma^2_{gbm}$: variância dos efeitos temporários de parcela ou da interação parcela dentro de locais x medições.

Assumindo que a correlação genotípica através das medições aproxima 1 (ou seja, assumindo que o caráter é o mesmo de uma colheita para outra), o modelo se reduz a $Y_{ijkn} = \beta_{jkn} + g_i + gl_{in} + gb_{ij/n} + gbml_{ijk/n} = \beta + g_p + g_{pn} + p_p + p_t$, o qual é denominado modelo de repetibilidade e interação genótipo x locais. Modelos desse tipo foram implementados nos modelos 69, 71, 151 e 152 do *software* Selegen-Reml/Blup.

5 SELEÇÃO COM MEDIDAS REPETIDAS

Quanto à seleção, algumas opções práticas existem: (a) atribuir pesos ou importâncias iguais para todas as safras (isto está implícito no modelo univariado de repetibilidade, ajustado quando as suposições são satisfeitas); (b) atribuir pesos diferentes às diferentes safras ou caracteres (isto é o que normalmente se faz na seleção para características múltiplas); (c) selecionar pelos valores genéticos e estabilidade através das safras via MHVG; (d) selecionar pelos valores genéticos e adaptabilidade (capacidade de melhorar em resposta à melhoria do ambiente) através das safras via PRVG; (e) selecionar conjuntamente pelos valores genéticos, estabilidade e adaptabilidade através das safras via MHPRVG.

Quando os valores genéticos preditos são obtidos para cada genótipo em cada safra (via g + gm, modelo multivariado, ARH ou SAD), há a possibilidade de aplicação de qualquer das cinco opções. A opção (a) equivale a obter a média dos valores genéticos preditos através das safras, selecionando então pelo valor genético médio ou por g no modelo g + gm. As opções (c), (d) e (e) também dão, implicitamente, importâncias iguais para as várias safras, embora conceitos adicionais (estabilidade e adaptabilidade) sejam simultaneamente considerados. Os métodos MHVG, PRVG e MHPRVG são apresentados no Capítulo 8.

Por outro lado, a opção (b) permite considerar a alteração do caráter com a idade e o sistema de utilização da cultura. Por exemplo, em cana-de-açúcar, a utilização da cultura se dá através de um corte em cana-planta e vários cortes em cana-soca. Assim, provavelmente, deva-se dar maiores pesos aos valores genéticos nas safras em cana-soca do que aos valores genéticos da safra em cana-planta. O mesmo raciocínio é válido para erva-mate e fruteiras, em que a produção por planta vai se estabilizando com a idade. Neste caso, as últimas safras poderiam receber maior peso. Em forrageiras, safras das águas e das secas poderiam receber diferentes pesos de acordo com a região de plantio. Também em forrageiras, a seleção por MHPRVG poderá ser relevante. Em seringueira, a produção de látex segue um regime anual com picos e decréscimos durante o ano, associado ao padrão de florescimento e desenvolvimento das sementes, sendo as maiores

produções verificadas quando as plantas estão livres das cargas de florescimento e desenvolvimento de frutos. Assim, a seleção por MHPRVG (estabilidade ao longo do ano) certamente será relevante também para seringueira.

CAPÍTULO 10
ANÁLISE ESTATÍSTICA EM SILVICULTURA, FRUTICULTURA, FORRAGICULTURA, AGRICULTURA E OLERICULTURA

Nesse capítulo são apresentadas algumas aplicações típicas em várias categorias de plantas, por meio do uso do software Selegen-Reml/Blup. Logicamente, as aplicações apresentadas referem-se apenas a alguns modelos pertinentes. Muitos outros modelos e situações experimentais são também importantes, mas seria impossível abordar todos eles.

1 ESPÉCIES FLORESTAIS

1.1 Estimação de componentes de variância e seleção em teca (*Tectona grandis*) em múltiplos experimentos na Costa Rica

Neste tópico são descritas algumas aplicações em espécies florestais. Como exemplo são apresentados resultados do programa de melhoramento de teca na Costa Rica, conduzido por uma cooperativa de melhoramento genético florestal, composta por doze empresas membros. Esses resultados foram relatados por Murillo et al. (2007). Os aspectos relatados aqui são igualmente aplicáveis a palmáceas cultivadas para produção de palmito tais quais pupunha, açaí, palmeira real e juçara. Aplicam-se também a espécies frutíferas quando avaliadas em uma só colheita.

A rede experimental aqui analisada consta de quatro experimentos instalados no delineamento experimental de blocos ao acaso com seis repetições, 28 famílias de meios irmãos e três plantas por parcela. As análises individuais por experimento foram realizadas empregando-se o modelo 1 do *software* Selegen-Reml/Blup.

O modelo estatístico equivale a y = Xr + Za + Wp + e, em que y é o vetor de dados, r é o vetor dos efeitos de repetição (assumidos como fixos) somados à média geral, a é o vetor dos efeitos genéticos aditivos individuais (assumidos como aleatórios), p é o vetor dos efeitos de parcela, e é o vetor de erros ou resíduos (aleatórios). As letras maiúsculas representam as matrizes de incidência para os referidos efeitos. Os resultados para a variável diâmetro são apresentados na Tabela 40.

Tabela 40. Estimativas de parâmetros genéticos para o caráter diâmetro em teca (T*ectona grandis*) cultivada na Costa Rica em quatro sítios. Análises de experimentos individuais.

Estimativas	Sítio 1 – AR	Sítio 2 - EC	Sítio 3 – JI	Sítio 4 – LE
Va	2.2193	1.5151	3.5156	0.4894
Vparc	0.0332	0.1296	0.3953	0.0517
Ve	3.3979	2.3569	7.3292	5.5953
Vf	5.6504	4.0016	11.2401	6.1364
h2a	0.3928	0.3786	0.3128	0.0798
h2aj	0.3951	0.3913	0.3242	0.0804
c2parc	0.0059	0.0324	0.0352	0.0084
h2mp	0.6592	0.6372	0.5865	0.2647
Acprog	0.8119	0.7982	0.7659	0.5145
h2ad	0.3288	0.3253	0.2646	0.0616
CVgi%	8.9359	10.5062	10.8558	4.2434
CVgp%	4.4679	5.2531	5.4279	2.1217
Cve%	7.8684	9.7093	11.1628	8.6619
CVr	0.5678	0.5410	0.4862	0.2449
PEV	0.1890	0.1374	0.3634	0.0899
SEP	0.4348	0.3707	0.6028	0.2999
Média	16.6712	11.7160	17.2717	16.4857

Os parâmetros apresentados na Tabela 40 são definidos a seguir.

Va: variância genética aditiva.

Vparc: variância ambiental entre parcelas.

Ve: variância residual (ambiental + não aditiva).

Vf = Va + Vparc + Ve: variância fenotípica individual.

h2a = Va/Vf: herdabilidade individual no sentido restrito, ou seja, dos efeitos aditivos.

h2aj = Va/(Va + Ve): herdabilidade individual no sentido restrito, ajustada para os efeitos de parcela.

c2parc = Vparc/Vf: coeficiente de determinação dos efeitos de parcela.

h2mp = (0.25 Va)/ [0.25 Va + Vparc/6 + (0.75 Va + Ve)/18]: herdabilidade da média de progênies, assumindo sobrevivência completa, em que 6 é o número de repetições e 18 é o número de repetições multiplicado pelo número de plantas por parcela (3).

Acprog = raiz quadrada de h2mp: acurácia da seleção de progênies, assumindo sobrevivência completa.

h2ad = (0.75 Va)/ (0.75 Va + Ve): herdabilidade aditiva dentro de parcela.

CVgi% = $[(Va)^{1/2}$ / Media Geral] * 100: coeficiente de variação genética aditiva individual ou ***evolvabilidade***.

CVgp%: $[(Va/4)^{1/2}$ / Media Geral] * 100: coeficiente de variação genotípica entre progênies.

CVe% = $\{[(0.75 Va + Ve)/3 + Vparc]^{1/2}$ / Media Geral} * 100: coeficiente de variação experimental.

CVr = CVgp/CVe = coeficiente de variação relativa.

PEV = $(1 – Acprog^2)$ Va/4: variância do erro de predição dos valores genotípicos de progênie, assumindo sobrevivência completa.

SEP = raiz quadrada da PEV: desvio padrão do valor genotípico predito de progênie, assumindo sobrevivência completa.

Média geral do experimento.

Pelos resultados da Tabela 40, constata-se a presença de considerável variabilidade genética para o caráter diâmetro nos três primeiros sítios, conforme pode ser visto pelas magnitudes do coeficiente de variação genética individual (CVgi%), os quais variaram de 9 % a 11 %. Tal variabilidade genética associada aos baixos valores do coeficiente de variação experimental (CVe%) propiciaram herdabilidades individuais de alta magnitude, variando de 31 % a 39 %. Estes valores revelam uma situação muito favorável para a seleção e ótimas perspectivas para o programa de

melhoramento genético da teca na Costa Rica. Por outro lado, o sítio 4 propiciou menor expressão de variabilidade genética e herdabilidade individual baixa, da ordem de 8 %. Em Genética Quantitativa, coeficientes de variação genética são denominados ***evolvabilidade*** (Houle, 1992), por estarem relacionados ao processo evolutivo ou capacidade de evolver dos organismos. A evolvabilidade expressa a quantidade de variação genética proporcional à média do caráter e tem-se mostrado aproximadamente constante para um caráter, mesmo após processos seletivos que alteram a média desse caráter.

Os coeficientes de variação relativa (CVr) nos três primeiros sítios foram maiores que 0.45 e, associados ao bom número de repetições, conduziram a altas (59 % a 66 %) confiabilidades ou herdabilidades ao nível de médias de famílias e altas (77 % a 81 %) acurácias seletivas. Conseqüentemente, a precisão da seleção nesses sítios será alta, conforme corroborado pelos baixos desvios padrões dos erros de predição (SEP). Por outro lado, a acurácia seletiva da seleção de famílias no sítio 4 foi menor (51 %). Nas etapas iniciais e intermediárias do melhoramento, acurácias da ordem de 70 % ou mais são desejáveis. Em termos de crescimento médio das plantas, os sítios 1, 3 e 4 apresentaram crescimento similar, enquanto o sítio 2 apresentou crescimento bem inferior.

As herdabilidades para a seleção dentro de famílias (h2ad) apresentaram magnitudes similares àquelas das herdabilidades individuais nos quatro sítios. Os coeficientes de determinação dos efeitos de parcela (c2parc) foram de baixa magnitude em todos os sítios, revelando que o delineamento experimental empregado foi ótimo pois não permaneceu heterogeneidade ambiental dentro de blocos.

As análises conjuntas dos experimentos dois a dois e também conjunta dos quatro experimentos foram realizadas empregando-se os modelos 4 e 51 do *software* Selegen-Reml/Blup. O modelo estatístico é dado por y = Xr + Za + Wp + Ti + e, em que y é o vetor de dados, r é o vetor dos efeitos de repetição (assumidos como fixos) somados à média geral, a é o vetor dos efeitos genéticos aditivos individuais (assumidos como aleatórios), p é o vetor dos efeitos de parcela (assumidos como aleatórios), i é vetor dos efeitos da interação genótipo x ambiente (aleatórios) e e é o vetor de erros ou resíduos (aleatórios). As letras maiúsculas representam as matrizes de incidência para os referidos efeitos. O vetor r contempla todas as repetições de todos os locais

(ajusta combinações repetição-local). Nesse caso, esse vetor contempla os efeitos de locais e de repetições dentro de locais. Os resultados para a variável diâmetro são apresentados na Tabela 41.

Tabela 41. Estimativas de parâmetros genéticos e estudo da interação genótipos x ambientes para o caráter diâmetro em teca (T*ectona grandis*) cultivada na Costa Rica em quatro sítios. Análises conjuntas de experimentos.

Estimativas	Sítios 1-2	Sítios 1-3	Sítios 1-4	Sítios 2-3	Sítios 2-4	Sítios 3-4	Sítios 1-2-3-4
Va	1.6128	2.1924	1.8975	1.2896	0.6042	1.7600	1.5943
Vparc	0.0703	0.1738	0.0657	0.2526	0.1772	0.3070	0.1592
Vint	0.0569	0.1909	0.0115	0.3104	0.2381	0.2117	0.1813
Ve	2.8878	6.2893	5.2032	5.4108	4.6822	8.0329	5.2938
Vf	4.6278	8.8464	7.1779	7.2635	5.7016	10.3117	7.2286
h2a	0.3485	0.2478	0.2643	0.1776	0.1060	0.1707	0.2206
c2parc	0.0152	0.0196	0.0092	0.0348	0.0311	0.0298	0.0220
c2int	0.0123	0.0216	0.0016	0.0427	0.0418	0.0205	0.0251
h2mp	0.7313	0.6240	0.7082	0.4771	0.3515	0.5281	0.7370
Acprog	0.8552	0.7899	0.8416	0.6907	0.5929	0.7267	0.8585
h2ad	0.2952	0.2073	0.2148	0.1517	0.0882	0.1411	0.1843
rgloc	0.8763	0.7417	0.9762	0.5095	0.3881	0.6751	0.6874
PEV	0.1083	0.2061	0.1384	0.1686	0.0979	0.2076	0.0464
SEP	0.3291	0.4540	0.3720	0.4106	0.3130	0.4557	0.2151
Média	14.1929	16.9665	17.8903	14.4937	15.4277	18.1940	16.1888

Os parâmetros apresentados na Tabela 41 e que não estão presentes na Tabela 40 são definidos a seguir.

Vint: variância da interação genótipos x ambientes.

c2int = Vint / Vf: coeficiente de determinação dos efeitos da interação genótipos x ambientes.

rgloc = Va / (Va + 4Vint): correlação genotípica entre o desempenho das progênies nos vários ambientes.

Pelos resultados da Tabela 41, constata-se a presença de considerável variabilidade genética para o caráter diâmetro nas análises conjuntas. As herdabilidades individuais apresentaram magnitudes moderadas a altas em todas as combinações de sítio, exceto na análise conjunta dos sítios 2 e 4 que apresentou baixa herdabilidade (10 %). A análise conjunta envolvendo os quatro sítios apresentou herdabilidade individual de 22 %. Este valor revela uma situação muito favorável para a seleção e indica que uma única população de melhoramento poderá atender toda a região de plantio na Costa Rica. Isto foi corroborado pelos resultados referentes à interação genótipos x ambientes, discutidos a seguir.

Os coeficientes de determinação da interação genótipos x ambientes (c2int) revelaram que a interação explicou pequena proporção da variabilidade fenotípica total para todas as combinações de sítio, exceto as combinações 2-3 e 2-4, em que os valores de c2int explicaram acima de 4 % da variação total. Esses resultados de interação conduziram aos seguintes valores de correlação genotípica através dos ambientes (correlação tipo B ou rgloc): 0.88, 0.74, 0.97, 0.51, 0.39, 0.68 e 0.69 para as combinações de sítios 1-2, 1-3, 1-4, 2-3, 2-4, 3-4 e 1-2-3-4, respectivamente. Todas as correlações apresentaram alta magnitude exceto para as combinações do sítio 2 com os locais 3 e 4. Esse sítio 2 é menos produtivo e interagiu mais do que os demais. No entanto, a correlação envolvendo os quatro sítios em conjunto foi de alta magnitude, indicando que a interação não é problemática para o melhorista pois não é de natureza complexa, conforme conceito apresentado por Vencovsky e Barriga (1992). Assim, a situação é muito favorável para se conduzir o melhoramento da teca. Adicionalmente, verifica-se que o local 1 apresenta altas correlações com todos os demais sítios, inclusive com o sítio 2. Assim, o sítio 1 poderá ser escolhido como um local ideal para a condução dos ensaios componentes do programa de melhoramento da teca, visando à seleção de genótipos para atender todos os sítios de plantio.

As confiabilidades ou herdabilidades ao nível de médias de famílias e acurácias seletivas das análises conjuntas foram altas, exceto para a análise conjunta do local 2 com o 4. Para a análise conjunta envolvendo os 4 locais, esses valores foram de 74 % para a confiabilidade e 86 % para a acurácia. Esses valores são excelentes e comprovam a eficiência de uma seleção geral envolvendo os quatro experimentos.

A significância dos efeitos genéticos aditivos e dos efeitos da interação famílias x ambiente na análise conjunta envolvendo os quatro sítios foi avaliada pela análise de deviance (ANADEV) apresentada na Tabela 42.

Tabela 42. Análise de deviance (ANADEV) para o caráter diâmetro avaliado em quatro sítios na Costa Rica.

Efeito	Deviance	LRT (Qui-quadrado)	Comp.Var.	Coef. Determ.
Famílias	7701.07^{+}	23.46**	Va = 1.5943**	h2a = 0.2206**
Parcela	7679.80^{+}	2.19^{ns}	Vparc=0.1592^{ns}	c2parc = 0.0220^{ns}
Famílias X Sítios	7685.98^{+}	8.37**	Vint=0.1813**	c2int = 0.0251**
Resíduo	-	-	Ve = 5.2938	c2res = 0.7323
Modelo Completo	7677.61	-	-	c2total = 1.00

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5% e 1%, respectivamente
+ Deviance do modelo ajustado sem os referidos efeitos

Verifica-se que os efeitos genéticos e da interação genótipos x ambientes foram significativos e os efeitos de parcela não foram. Como a correlação genética através dos ambientes foi alta, constata-se que a interação, embora significativa, é de natureza simples, ou seja, devida à mudança de variabilidade genética de um ambiente para outro.

Na Tabela 43 são apresentados os resultados referentes à seleção das 10 melhores famílias para o caráter diâmetro segundo diferentes critérios de seleção conforme obtido pelo modelo 51.

Tabela 43. Melhores famílias de teca selecionadas em cada sítio, em todos os sítios e também com base na estabilidade genotípica, adaptabilidade genotípica e simultaneamente por produtividade, estabilidade e adaptabilidade.

Ordem	Sítio 1	Sítio 2	Sítio 3	Sítio 4	Todos os Sítios	Estabilidade (MHVG)	Adaptabilidade (PRVG)	Produtividade, Estabilidade e Adaptabilidade (MHPRVG)
1	**4**	**4**	**4**	**10**	4	4	4	4
2	**13**	**15**	**9**	**21**	13	13	13	13
3	**3**	**3**	**13**	**4**	9	9	9	9
4	**7**	**9**	**10**	**13**	10	3	3	10
5	**9**	20	2	**19**	3	15	10	3
6	**14**	**21**	**7**	**3**	15	10	15	15
7	**10**	**13**	**14**	**14**	7	7	7	7
8	**15**	**7**	**21**	**15**	14	21	14	14
9	**21**	**14**	**15**	5	21	14	21	21
10	**19**	23	**19**	**9**	19	19	19	19

Verifica-se que, dentre as dez melhores famílias na análise conjunta de todos os sítios, 10, 8, 9 e 9 estão dentre as dez melhores nos sítios 1, 2, 3 e 4, respectivamente. Ou seja, selecionando as melhores na análise global de todos os sítios, atende-se muito bem cada sítio em particular. Isto está de acordo com a alta correlação genotípica entre sítios, verificada na análise conjunta.

Outros ordenamentos foram obtidos explorando os conceitos de estabilidade (pequena variação na performance genotípica através dos ambientes), adaptabilidade (capacidade de resposta à melhoria do ambiente) e esses dois atributos simultaneamente e também considerando a produtividade genotípica média através dos ambientes. Para tanto, foram obtidas a média harmônica dos valores genéticos preditos (MHVG), a performance relativa dos valores genéticos preditos em relação à média de cada ambiente (PRVG) e a média harmônica da performance relativa dos valores genotípicos preditos (MHPRVG). Verifica-se pela Tabela 43 que, embora a ordem de seleção não seja exatamente a mesma, as dez melhores famílias por esses critérios coincidem com as dez melhores selecionadas com base na análise conjunta de todos os locais. Isto é muito bom pois revela que as famílias mais produtivas são também estáveis e de grande adaptabilidade.

Na Tabela 44 são apresentados os resultados da seleção de indivíduos para a composição de um pomar clonal de sementes, com base na análise conjunta de todos os locais usando o modelo 4 do Selegen Reml/Blup.

Tabela 44. Seleção dos 20 melhores indivíduos de teca nos quatro sítios na Costa Rica, para o caráter diâmetro, com base nos efeitos aditivos (a), valores genéticos aditivos (u + a), efeitos de dominância (d) e efeitos genotípicos totais (g). Sem restrição ao tamanho efetivo populacional (Ne)

Ordem	Bloco	Família	Árvore	f	a	u+a	Ganho	Média Melhorada	Ne	d	g
1	15	13	392	29.30	3.01	19.20	3.01	19.20	1.00	1.56	4.57
2	34	10	1809	28.70	2.40	18.59	2.71	18.90	2.00	1.21	3.62
3	34	13	1824	28.00	2.26	18.45	2.56	18.75	2.48	1.05	3.31
4	36	4	2028	25.20	2.12	18.31	2.45	18.64	3.49	0.77	2.89
5	31	4	1438	24.00	1.96	18.15	2.35	18.54	4.11	0.67	2.63
6	4I	10	2169	28.60	1.93	18.12	2.28	18.47	4.80	0.90	2.83
7	34	7	1797	26.00	1.82	18.01	2.22	18.41	5.72	0.91	2.74
8	23	20	914	21.50	1.81	17.99	2.17	18.35	6.68	1.08	2.89
9	34	13	1822	25.50	1.80	17.99	2.12	18.31	6.94	0.74	2.54
10	4II	3	2229	27.70	1.77	17.96	2.09	18.28	7.89	0.83	2.61
11	36	10	2059	24.50	1.74	17.93	2.06	18.25	8.21	0.77	2.51
12	4II	4	2233	25.20	1.72	17.91	2.03	18.22	8.57	0.50	2.22
13	34	13	1823	24.90	1.69	17.88	2.00	18.19	8.69	0.67	2.36
14	13	4	185	21.80	1.69	17.88	1.98	18.17	8.88	0.48	2.17
15	34	15	1836	24.60	1.66	17.85	1.96	18.15	9.77	0.80	2.46
16	23	4	845	16.80	1.64	17.83	1.94	18.13	9.76	0.45	2.10
17	32	8	1573	24.70	1.60	17.79	1.92	18.11	10.64	0.96	2.57
18	31	13	1486	22.80	1.59	17.78	1.90	18.09	10.68	0.61	2.20
19	31	3	1435	22.50	1.59	17.77	1.88	18.07	11.36	0.71	2.29
20	26	4	1293	16.00	1.58	17.77	1.87	18.06	11.27	0.41	2.00

Constata-se que o melhor indivíduo para o caráter diâmetro, dentre todos avaliados nos quatro sítios, pertence à família 15 e encontra-se no bloco 15, ou seja, bloco 5 do local 1. Seu valor fenotípico observado (f) é de 29.3 cm e seu valor genético aditivo é de 19.20. Assim, em uma reprodução via sementes desse indivíduo, metade desse valor genético aditivo será transmitido para a sua descendência. O efeito genotípico (g = 4.57) desse indivíduo somado à média geral (16.19, conforme a Tabela 41) fornece o valor genotípico ou valor clonal do mesmo, que no caso é de 20.76. Esse valor foi obtido assumindo dominância alélica completa em uma população com nível intermediário de melhoramento.

A seleção dos 20 melhores indivíduos para estabelecimento de um pomar de sementes conduzirá a um ganho genético de 1.87 cm sobre a média geral (16.19) e a média da população melhorada em uma próxima geração de plantio será de 18.06. O ganho genético será então de 11,6 %. Esses 20 indivíduos selecionados estão associados a um tamanho efetivo ou genético de população igual a 11.27. Esse se refere ao número equivalente em termos de indivíduos não aparentados. O número efetivo (11) é menor do que o número físico (20) porque vários desses indivíduos são meios irmãos devido ao fato de pertecerem à mesma família. Por exemplo, a família 13 contribuiu com cinco indivíduos. Com esse Ne, o coeficiente de endogamia ou de endocruzamento associado às sementes produzidas no pomar é de F = [1/(2Ne)] = 4,5 %. Uma otimização da seleção com restrição no Ne pode reduzir esse coeficiente de endogamia.

Na Tabela 45 são apresentados os resultados da seleção de indivíduos para a composição de um pomar clonal de sementes, após otimização da seleção com restrição na endogamia e no tamanho efetivo populacional (Ne) usando o modelo 106 do Selegen Reml/Blup.

Tabela 45. Seleção dos 20 melhores indivíduos de teca nos quatro sítios na Costa Rica, para o caráter diâmetro, com base nos efeitos aditivos (a) e valores genéticos aditivos (u + a), com restrição ao tamanho efetivo populacional (Ne).

Ordem	Bloco	Família	Árvore	f	a	u+a	Ganho	Media Melhorada	Ne
1	15	13	392	29.30	3.01	19.20	3.01	19.20	1.00
2	34	10	1809	28.70	2.40	18.59	2.71	18.90	2.00
3	34	13	1824	28.00	2.26	18.45	2.56	18.75	2.48
4	36	4	2028	25.20	2.12	18.31	2.45	18.64	3.49
5	31	4	1438	24.00	1.96	18.15	2.35	18.54	4.11
6	4I	10	2169	28.60	1.93	18.12	2.28	18.47	4.80
7	34	7	1797	26.00	1.82	18.01	2.22	18.41	5.72
8	23	20	914	21.50	1.81	17.99	2.17	18.35	6.68
9	4II	3	2229	27.70	1.77	17.96	2.12	18.31	7.66
10	34	15	1836	24.60	1.66	17.85	2.08	18.26	8.64
11	32	8	1573	24.70	1.60	17.79	2.03	18.22	9.63
12	31	3	1435	22.50	1.59	17.77	2.00	18.18	10.23
13	4I	5	2154	28.70	1.55	17.74	1.96	18.15	11.21
14	14	7	288	25.60	1.54	17.73	1.93	18.12	11.83
15	33	9	1678	23.60	1.47	17.66	1.90	18.09	12.81
16	4I	19	2187	27.10	1.45	17.64	1.87	18.06	13.79
17	15	21	413	21.40	1.41	17.60	1.85	18.03	14.77
18	4III	21	2351	24.60	1.41	17.60	1.82	18.01	15.38
19	12	9	114	22.00	1.41	17.60	1.80	17.99	16.00
20	11	14	42	22.20	1.38	17.57	1.78	17.97	16.97

Com a restrição admitindo no máximo dois indivíduos selecionados por família, o ganho genético equivaleu a 1.78 cm e a média da população melhorada será de 17.97 (Tabela 45), ou seja, o ganho será de 11 %. Esse ganho é praticamente igual ao obtido na seleção sem restrição (Tabela 44). No entanto, o Ne aumentou de 11 para 17, um aumento de 54,5 %. O coeficiente de endogamia ou de endocruzamento associado às sementes produzidas no pomar agora é de F = 2,9 %. Assim, essa seleção com restrição no Ne deve ser praticada para estabelecimento do pomar de produção de sementes comerciais.

Para estabelecimento do pomar é relevante verificar se os indivíduos do experimento são de fato superiores às suas matrizes. Na Tabela 46 são apresentados resultados referentes à seleção com sobreposição de gerações.

Tabela 46. Seleção dos 20 melhores indivíduos (matrizes e progênies) de teca nos quatro sítios na Costa Rica, para o caráter diâmetro, com base nos efeitos aditivos (a) e com sobreposição de gerações, ou seja, incluindo matrizes e progênies em um mesmo ordenamento.

Ordem	Bloco	Família	Árvore	a
1	15	13	392	3.01
2	34	10	1809	2.40
3	34	13	1824	2.26
4	36	4	2028	2.12
5	31	4	1438	1.96
6	4I	10	2169	1.93
7	**0**	**4**	**0**	**1.93**
8	34	7	1797	1.82
9	23	20	914	1.81
10	34	13	1822	1.80
11	4II	3	2229	1.77
12	36	10	2059	1.74
13	4II	4	2233	1.72
14	34	13	1823	1.69
15	13	4	185	1.69
16	34	15	1836	1.66
17	23	4	845	1.64
18	32	8	1573	1.60
19	31	13	1486	1.59
20	31	3	1435	1.59

Verifica-se que, dentre os 20 melhores indivíduos no geral, entre progênies e matrizes, apenas um deles é uma matriz da geração anterior. Essa matriz (4) é o sétimo indivíduo no ordenamento e aparece com o número 0 na coluna de blocos. No entanto, tal matriz apresenta dois filhos melhores do que ela, o quarto e quinto indivíduos do ordenamento. Assim, tal matriz não deve ser incluída no pomar de sementes mas apenas os seus dois filhos. Então, o pomar será formado apenas por indivíduos das progênies, ou seja, não conterá qualquer matriz.

Na Tabela 47 é apresentado o resultado do agrupamento de famílias com base nas distâncias genéticas euclidianas quadráticas obtidas a partir de análise multivariada (modelo 104) envolvendo vários caracteres.

Tabela 47. Agrupamento genético das famílias com base na distância Euclideana quadrática e método multivariado mutuamente exclusivo de Tocher.

Grupos Genéticos	Famílias
1	3 6 7 8 10 12 13 14 19 21 23 25 27 28 33
2	2 5 24 29
3	11 16 22
4	9 15
5	1 35
6	20
7	4

Verifica-se a formação de sete grupos distintos de famílias, sendo que o grupo maior contempla 15 famílias. O cruzamento entre indivíduos selecionados pertencentes a excelentes famílias de diferentes grupos deve ser enfatizado visando aumentar a probabilidade de obtenção de alta capacidade específica de combinação ou heterose. Por exemplo, os melhores indivíduos das

famílias 4 (grupo 7) e 20 (grupo 6) devem ser cruzados com os melhores indivíduos das famílias 10 e 13, as quais pertencem ao grupo 1. Assim, famílias excepcionais de irmãos completos poderão ser geradas, visando à seleção de clones heteróticos dentro das mesmas.

Os grupos acima podem também direcionar a composição de duas sublinhas de melhoramento ou mesmo duas populações para a condução da seleção recorrente recíproca visando ao melhoramento complementar (uma em função da outra) das mesmas, ou seja, melhoramento do cruzamento entre elas. Uma população poderia ser formada a partir da seleção de indivíduos de famílias do grupo 1 e outra poderia ser formada pela seleção de indivíduos de famílias dos demais grupos.

Os procedimentos apresentados no tópico seguinte são também de grande aplicação em programas de melhoramento florestal, especialmente no caso da seleção recorrente recíproca aplicada em eucalipto e pínus. Nessa situação, cruzamentos fatoriais interpopulacionais são muito usados.

2 ESPÉCIES FRUTEIRAS E FRUTÍFERAS

Neste tópico são descritas algumas aplicações em fruticultura e outras culturas frutíferas agroindustriais tais como palmáceas (açaí, pupunha, dendê, coco, tâmara, macaúba, buriti, babaçú, butiá, gariroba), plantas estimulantes (café, cacau, guaraná) e energéticas (pinhão manso). Os procedimentos apresentados são aplicáveis também nas espécies florestais erva-mate e seringueira, as quais não são frutíferas em termos de uso comercial mas são avaliadas em nível individual e com medidas repetidas em cada indivíduo. Esses procedimentos e também aqueles apresentados no tópico 10.1 são aplicáveis também na domesticação e melhoramento de várias espécies frutíferas nativas de importância na América Latina, conforme a Tabela 48.

Tabela 48. Espécies frutíferas nativas sob domesticação e/ou melhoramento.

Espécie	Nome Científico	Bioma de Origem
Açaí	*Euterpe oleracea*	Amazônia
Buriti	*Mauritia flexuosa*	Amazônia e Cerrado
Camu-camu	*Myrciaria dubia*	Amazônia
Castanha-do-Pará	*Bertholetia excelsa*	Amazônia
Cupuaçu	*Thebroma grandiflorum*	Amazônia
Cherimóia	*Annona cherimoia*	Amazônia
Graviola	*Annona muricata*	Amazônia
Pinha	*Annona squamosa*	Amazônia
Pupunha	*Bactris gasipaes*	Amazônia
Babaçu	*Orbignya speciosa*	Caatinga
Cajá	*Spondias lutea*	Caatinga
Ciriguela	*Spondias purpurea*	Caatinga
Sapoti	*Manilkara sapota*	Caatinga
Umbu	*Spondias tuberosa*	Caatinga
Araçá	*Psidium rufum*	Cerrado
Barú – Castanha-do-Cerrado	*Dipteryx alata*	Cerrado
Cagaita	*Eugenia dysenterica*	Cerrado
Cajuzinho	*Anacardium humile*	Cerrado
Gabiroba	*Campomanesia guabiroba*	Cerrado
Gariroba – Palmito-amargo	*Syagrus oleracea*	Cerrado
Jenipapo	*Genipa americana*	Cerrado e Amazônia
Macaúba	*Acrocomia aculeata*	Cerrado
Mangaba	*Hancornia speciosa*	Cerrado e Caatinga
Pequi	*Caryocar brasiliense*	Cerrado
Pera-do-Cerrado	*Eugenia klotzschiana*	Cerrado
Pitanga	*Eugenia uniflora*	Cerrado
Butiá	*Butia capitata*	Sul
Goiabeira-serrana	*Feijoa sellowiana*	Sul
Pinhão	*Araucária angustifolia*	Sul
Uvalha	*Eugenia uvalha*	Sul
Jaboticaba	*Myrciaria cauliflora*	Mata Atlântica
Sapucaia	*Lecythis pisonis*	Mata Atlântica e Amazônia

2.1 Análise de cruzamentos fatoriais interpopulacionais com medidas repetidas em cajueiro

Como exemplo será considerada a avaliação de progênies de irmãos germanos obtidas sob cruzamentos fatoriais interpopulacionais no programa de melhoramento de caju da *Embrapa Agroindústria Tropical* localizada em Fortaleza, CE. Esses resultados foram publicados por Cavalcanti et al. (2007). Os aspectos descritos aqui são igualmente aplicáveis nas demais espécies mencionadas nesse tópico.

O experimento constou da avaliação de 20 famílias de irmãos completos obtidas sob cruzamentos em esquema fatorial envolvendo uma população de cajueiro anão (quatro genitores) e outra de cajueiro comum (cinco genitores). O delineamento experimental foi o de blocos ao acaso com três repetições e cinco plantas por parcela. Foram avaliadas as variáveis altura das plantas em metros no quarto ano e produção de castanha em gramas por planta em quatro colheitas consecutivas anuais.

Inicialmente foram realizadas análises para cada variável empregando-se o modelo 88 do *software* Selegen-Reml/Blup, descrito a seguir.

$$y = Xb + Z_m g_m + Z_f g_f + Ws + Tp + \varepsilon$$

em que:

y, b, g_m, g_f, s, p e ε: vetores dos dados observados, vetor dos efeitos de bloco (assumidos como fixos), vetor dos efeitos da capacidade geral de combinação dos genitores na população utilizada como masculina (assumidos como aleatórios), vetor dos efeitos da capacidade geral de combinação dos genitores na população utilizada como feminina (assumidos como aleatórios), vetor dos efeitos da capacidade específica de combinação entre os genitores das duas populações (assumidos como aleatórios), vetor dos efeitos de parcela (assumidos como aleatórios) e erros aleatórios, respectivamente.

X, Z_m, Z_f, W e T: matrizes de incidência para b, g_m, g_f, s e p, respectivamente.

As distribuições e estruturas de variância associadas aos termos do modelo são:

$$y|b,V \sim N(Xb,\ V)$$
$$g_m|I\sigma^2_{g_m} \sim N(0, I\sigma^2_{g_m})$$
$$g_f|I\sigma^2_{g_f} \sim N(0, I\sigma^2_{g_f})$$
$$s|\sigma^2_s \sim N(0, I\sigma^2_s)$$
$$p|\sigma^2_p \sim N(0, I\sigma^2_p)$$
$$\varepsilon|\sigma^2_\varepsilon \sim N(0, I\sigma^2_\varepsilon)$$
$$V = Z_m\sigma^2_{g_m}Z_m'+Z_f\sigma^2_{g_f}Z_f'+WI\sigma^2_sW'+TI\sigma^2_pT'+I\sigma^2_e$$

Sob esse modelo, os efeitos de g_m e g_f estão associados aos efeitos aditivos interpopulacionais nas duas populações. Os efeitos s referem-se aos efeitos de dominância associados ao cruzamento entre as duas populações. As estimativas da variância aditiva e de dominância e dos coeficientes de herdabilidade foram estimados por meio das seguintes fórmulas:

$\hat{\sigma}^2_a = 2(\hat{\sigma}^2_{g_m} + \hat{\sigma}^2_{g_f})$: estimativa da variância aditiva interpopulacional.

$\hat{\sigma}^2_d = \dfrac{4\hat{\sigma}^2_s}{\hat{\sigma}^2_y}$: estimativa da variância de dominância interpopulacional.

$\hat{\sigma}^2_y = \hat{\sigma}^2_{g_m} + \hat{\sigma}^2_{g_f} + \hat{\sigma}^2_s + \hat{\sigma}^2_p + \hat{\sigma}^2_\varepsilon$: estimativa da variância fenotípica interpopulacional.

$\hat{h}^2_a = \dfrac{\hat{\sigma}^2_a}{\hat{\sigma}^2_y}$: estimativa da herdabilidade individual no sentido restrito.

$\hat{h}^2_g = \dfrac{\hat{\sigma}^2_a + \hat{\sigma}^2_d}{\hat{\sigma}^2_y}$: estimativa da herdabilidade individual no sentido amplo.

$\hat{h}^2_d = \dfrac{\hat{\sigma}^2_d}{\hat{\sigma}^2_y}$ = estimativa do coeficiente de determinação individual dos efeitos de dominância.

$\hat{c}^2_{g_m} = \dfrac{\hat{\sigma}^2_{g_m}}{\hat{\sigma}^2_y}$: estimativa do coeficiente de determinação dos efeitos da capacidade geral de combinação na população usada como masculina.

$\hat{c}^2_{g_f} = \frac{\hat{\sigma}^2_{g_f}}{\hat{\sigma}^2_y}$: estimativa do coeficiente de determinação dos efeitos da capacidade geral de combinação na população usada como feminina.

$\hat{c}^2_s = \frac{\hat{\sigma}^2_s}{\hat{\sigma}^2_y}$: estimativa do coeficiente de determinação dos efeitos da capacidade específica de combinação interpopulacional.

$\hat{c}^2_p = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_y}$: estimativa do coeficiente de determinação dos efeitos de parcela.

$\hat{c}^2_p = \frac{(\hat{\sigma}^2_{g_m} + \hat{\sigma}^2_{g_f})}{\hat{\sigma}^2_y}$: estimativa do coeficiente de determinação dos efeitos da capacidade geral de combinação interpopulacional.

Os resultados para os caracteres altura e produção de castanha no quarto ano são apresentados na Tabela 49.

Tabela 49. Estimativas de parâmetros genéticos para os caracteres altura da planta e produção de castanhas em cajueiro. Análises individuais por ano.

Estimativas	Altura da Planta	Produção de Castanha
$\hat{\sigma}^2_{g_m}$	0.089	64558.850
$\hat{\sigma}^2_{g_f}$	0.161	53724.610
$\hat{\sigma}^2_a$	0.501	236566.910
$\hat{\sigma}^2_s$	0.004	60564.080
$\hat{\sigma}^2_p$	0.018	26868.590
$\hat{\sigma}^2_\varepsilon$	0.306	1448839.610
$\hat{\sigma}^2_y$	0.579	1654555.740
$\hat{c}^2_g$	0.432	0.071
$\hat{c}^2_s$	0.008	0.037
$\hat{h}^2_d$	**0.031**	**0.146**
$\hat{h}^2_a$	**0.865**	**0.143**
$\hat{h}^2_g$	**0.895**	**0.289**
$\hat{c}^2_p$	0.032	0.016
Média	3.096	2037.204

Constata-se que o caráter altura da planta em cajueiro apresenta controle genético predominantemente aditivo e é de alta herdabilidade (próxima a 90 %). A herdabilidade ou coeficiente de determinação dos efeitos de dominância foi praticamente zero para esse caráter. Esses resultados indicam que o melhoramento desse caráter pode ser realizado aplicando-se métodos mais simples de seleção e que a estratégia de melhoramento a ser adotada deve ser a seleção recorrente intrapopulacional a qual capitaliza os efeitos genéticos aditivos.

Por outro lado, o caráter produção de castanha apresenta baixa herdabilidade aditiva (14,3 %) e alta dominância alélica proporcionalmente aos efeitos aditivos. A herdabilidade ou coeficiente

de determinação dos efeitos de dominância (14,6 %) foi praticamente igual à herdabilidade aditiva. Esses resultados indicam que o melhoramento desse caráter deve ser realizado aplicando-se métodos mais elaborados de seleção e que a estratégia de melhoramento a ser adotada deve ser a seleção recorrente recíproca a qual melhora o híbrido interpopulacional, capitalizando tanto os efeitos genéticos aditivos quanto os de dominância.

Os caracteres altura e produção de castanha apresentam correlação genética positiva e alta. Isto dificulta a seleção de plantas produtivas e com porte baixo, mais próximas do tipo anão. Uma alternativa interessante é selecionar com base na variável relacional produção de castanha por unidade de altura (P/A). Para isto basta criar a nova variável P/A dividindo a coluna de dados da produção pela coluna de dados da altura. Expressando a produção dessa forma, a comparação entre a produtividade de plantas com diferentes tamanhos fica mais precisa e a seleção beneficiará aquelas concomitantemente mais produtivas e mais baixas. Outra alternativa é expressar a produção por unidade de área de copa, calculando-se a área da copa com base na altura (A) e diâmetro (D) da planta e assumindo uma forma esférica. Nesse caso, a área lateral de uma zona esférica é dada por πDA.

Realizou-se também uma análise simultânea dos dados de produção de castanha nas quatro colheitas. Utilizou-se o modelo 100 do Selegen-Reml/Blup, o qual permite estimar simultaneamente as herdabilidades nos sentidos restrito e amplo e também a repetibilidade do caráter, considerando toda a estrutura experimental e o esquema de cruzamentos fatoriais. Tal modelo é dado por: $y = X\beta + Z_m g_m + Z_f g_f + Ws + Tp + Qc + \varepsilon$ em que c é um vetor de efeitos de ambiente permanente ou efeitos ambientais constantes de uma colheita à outra associados a cada indivíduo. Esse efeito foi assumido como tendo distribuição $c|\sigma_c^2 \sim N(0, I\sigma_c^2)$.

O coeficiente de repetibilidade individual foi estimado por: $\hat{r} = \dfrac{\hat{\sigma}_{g_m}^2 + \hat{\sigma}_{g_f}^2 + \hat{\sigma}_s^2 + \hat{\sigma}_p^2 + \hat{\sigma}_c^2}{\hat{\sigma}_{g_m}^2 + \hat{\sigma}_{g_f}^2 + \hat{\sigma}_s^2 + \hat{\sigma}_p^2 + \hat{\sigma}_c^2 + \hat{\sigma}_\varepsilon^2}$.

Na Tabela 50 são apresentados os resultados obtidos com o uso desse modelo.

Tabela 50. Estimativas de parâmetros genéticos para o caráter produção de castanhas em cajueiro. Análise simultânea das quatro colheitas anuais.

Estimativas	**Produção de Castanha**
$\hat{\sigma}^2_{g_m}$	66961.300
$\hat{\sigma}^2_{g_f}$	2195.610
$\hat{\sigma}^2_a$	138313.820
$\hat{\sigma}^2_s$	45440.770
$\hat{\sigma}^2_p$	66278.180
$\hat{\sigma}^2_c$	450985.210
$\hat{\sigma}^2_\varepsilon$	853616.110
$\hat{\sigma}^2_y$	1034491.970
$\hat{c}^2_g$	0.067
$\hat{c}^2_s$	0.044
$\hat{h}^2_d$	**0.176**
$\hat{h}^2_a$	**0.134**
$\hat{h}^2_g$	**0.309**
$\hat{c}^2_p$	0.064
$\hat{c}^2_c$	0.202
$\hat{r}$	**0.284**
Média	2037.204

Verifica-se que a repetibilidade (0.284) e a herdabilidade no sentido amplo (0.309) apresentaram valores muito próximos, revelando que os efeitos de ambiente permanente são muito baixos. É importante relatar que o componente $\hat{c}^2_c$ refere-se aos efeitos permanentes de indivíduo dentro de família e inclui ambiente permanente mais metade da variação genética aditiva e mais três

quartos da variação genética de dominância. Teoricamente, a repetibilidade deve ser maior ou igual à herdabilidade no sentido amplo. No presente caso, tal herdabilidade estimada não atendeu a esse requisito. Isto revela que a multiplicação da variância da capacidade específica de combinação por 4, visando obter o coeficiente de determinação dos efeitos de dominância, é um procedimento apenas aproximado. Isto porque a variância devida aos efeitos genéticos epistáticos é também multiplicada por 4, fato que inflaciona a herdabilidade no sentido amplo.

A repetibilidade ao nível de média das quatro safras (ou coeficiente de determinação) equivaleu a 0.613 e a acurácia fenotípica permanente associada à seleção individual com base nas quatro safras foi 0.783. A Tabela 51 apresenta os números de medições necessários para se atingir determinada acurácia na seleção de indivíduos.

Tabela 51. Número de colheitas necessário para se atingir determinados valores de acurácia seletiva.

Número de Colheitas	Repetibilidade Individual	Coeficiente de Determinação Fenotípica Permanente	Acurácia Fenotípica Permanente	Herdabilidade Individual Aditiva em Nível de Média de Colheitas	Acurácia Genética Aditiva	Eficiência em Relação ao Uso de Apenas Uma Colheita
1	0.284	0.28	0.53	0.13	0.37	1.00
2	0.284	0.44	0.67	0.21	0.46	1.25
3	0.284	0.54	0.74	0.26	0.51	1.38
4	0.284	0.61	0.78	0.29	0.54	1.47
5	0.284	0.66	0.82	0.31	0.56	1.53
6	**0.284**	**0.70**	**0.84**	**0.33**	**0.58**	**1.58**
7	0.284	0.74	0.86	0.35	0.59	1.61
8	0.284	0.76	0.87	0.36	0.60	1.64
9	0.284	0.78	0.88	0.37	0.61	1.66
10	0.284	0.80	0.89	0.38	0.61	1.68

Verifica-se que, considerando a repetibilidade individual estimada, a adoção de seis colheitas conduz a 70 % de determinação fenotípica permanente, 84 % de acurácia fenotípica permanente e

58 % de acurácia genética aditiva. Com esse número de colheitas, a herdabilidade individual no sentido restrito passa de 13 % em nível de uma colheita para 33 % em nível das seis colheitas. A eficiência do uso de seis colheitas em relação ao uso de apenas uma é de 1.58 ou 58%. Aumentando-se para sete colheitas, a eficiência passa de 1.58 para 1.61. Esses 3 % a mais provavelmente não são compensatórios. Assim, recomenda-se a adoção de seis colheitas anuais.

Esse modelo de análise permite as seguintes modalidades de seleção de acordo com os objetivos: (i) seleção de genitores na população masculina para cruzamento e recombinação em um programa de seleção recorrente recíproca; (ii) seleção de genitores na população feminina para cruzamento e recombinação em um programa de seleção recorrente recíproca; (iii) seleção de famílias (ou cruzamentos) para direcionar a seleção de clones potenciais ou para plantios comerciais via sementes ou via clonagem; (iv) seleção de clones potenciais nas progênies híbridas; (v) seleção de genitores potenciais para programas de seleção recorrente. A seguir são apresentados alguns resultados referentes a seleção de cruzamentos para o caráter produção de castanhas.

Tabela 52. Valores genotípicos em gramas dos cinco melhores cruzamentos selecionados para o caráter produção de castanhas em caju.

Ordem	Cruzamento	Genitor Masculino	Genitor Feminino	Média Geral	Metade do Efeito Genético Aditivo do Genitor Masculino	Metade do Efeito Genético Aditivo do Genitor Feminino	Efeito de Dominância do Cruzamento	Valor Genotípico Total do Cruzamento
1	17	4	7	2014.28	212.19	12.88	139.74	2379.09
2	16	4	6	2014.28	212.19	6.34	126.17	2358.98
3	18	4	8	2014.28	212.19	-0.87	105.23	2330.83
4	5	2	5	2014.28	132.67	3.78	143.58	2294.31
5	15	4	5	2014.28	212.19	3.78	29.87	2260.12

O valor genotípico total apresentado na última coluna é dado pela soma das quatro colunas anteriores. Verifica-se que o melhor cruzamento (17) capitaliza o melhor efeito aditivo de cada população e a segunda melhor capacidade específica de combinação. A melhor capacidade

específica de combinação é dada pelo cruzamento 5. Esse cruzamento e também o 17 tem potencial para ser melhorado pela seleção recorrente recíproca individual, ou seja, cada genitor de cada cruzamento ser melhorado um em função do outro. A seleção do melhor cruzamento propicia um ganho genético de 18 %.

O modelo analisado emite também como resultados os valores genéticos aditivos (para a seleção de genitores potenciais) e genotípicos (para a seleção de clones potenciais) de cada indivíduo das progênies, os quais são usados para fins de seleção individual. A forma dos resultados é similar àquela apresentada na Tabela 44.

3 ESPÉCIES FORRAGEIRAS E CANA-DE-AÇÚCAR

Neste tópico são descritas algumas aplicações em forragicultura, as quais podem também ser estendidas para a cultura da cana-de-açúcar. As aplicações aqui relatadas foram apresentadas por Resende et al. (2007).

3.1 Avaliação de *Panicum maximum* em Vários Locais e em Várias Colheitas

O presente tópico trata da avaliação de uma rede experimental de *Panicum maximum* da *Embrapa Gado de Corte*, com 30 acessos, estabelecida em três locais (Acre, Bahia e Paraná), no delineamento de blocos ao acaso com três repetições e uma observação total por parcela. Na Bahia e no Paraná foram realizados seis e no Acre oito cortes ou colheitas por parcela, avaliando-se a característica produção total de matéria seca. A rede experimental é desbalanceada em termos de números de cortes e também devido à perda de algumas parcelas.

Quatro modalidades de avaliação genética podem ser realizadas com esse conjunto de dados.

Avaliação em um só local e em uma só colheita

Tomando-se apenas os dados do Paraná, análises individuais foram realizadas para cada colheita, segundo o modelo 20 do Selegen-Reml/Blup. Os principais resultados são apresentados a seguir.

(1) Análises de Colheitas Individuais no Paraná

Tabela 53. Resultados das análises de colheitas individuais no Paraná.

Colheita	Herdabilidade individual (h2g)	Variância genotípica (Vg)	Variância fenotípica (Vf)	Média Geral do Caráter	Correção 1 Heterogeneidade Sm/Si	Correção 2 Heterogeneidade hgi/hgm
1	0.3929	0.0883	0.2247	1.7127	2.0338	1.0658
2	0.4715	0.3296	0.6989	3.0152	1.1531	1.1676
3	0.3065	0.5336	1.7407	5.9290	0.7307	0.9414
4	0.4739	0.0679	0.1433	1.1067	2.5465	1.1706
5	**0.0760**	0.0217	0.2852	1.4713	**1.8051**	**0.4687**
6	0.3545	0.8803	2.4830	5.4545	0.6118	1.0124
Média	**0.3458**	**0.3202**	**0.9293**	**3.1149**	**1.4801**	**0.9717**

Verifica-se presença de heterogeneidade de variâncias genotípica e fenotípica e também de herdabilidades entre colheitas. As estimativas de herdabilidade variaram de 7, 6 % a 47,4 %. Assim, para uma análise conjunta de todas as safras, recomenda-se alguma transformação de dados. Uma transformação comumente usada (Visscher et al., 1992) baseia-se na padronização dos dados de cada local multiplicando-os por Sm/Si, em que Si é o desvio padrão fenotípico na colheita i e Sm é o desvio padrão fenotípico médio para todas as colheitas. Resende (2004) relata que uma transformação melhor baseia-se na multiplicação dos dados de cada local por hgi/hgm, em que hgi é raiz quadrada da herdabilidade na colheita i e hgm é a raiz quadrada da média das herdabilidades nas várias colheitas. Verifica-se, pelos resultados acima, que a correção hgi/hgm mostra-se bem mais coerente pois dá peso 0,4687 para os dados da colheita com herdabilidade 7,6 % e peso 1,17 para os dados da colheita com herdabilidade 47,4 %. Essa ponderação é desejável visto que na predição dos valores genotípicos pelo BLUP na análise conjunta, uma única herdabilidade média é

usada para ponderar os dados de todas as colheitas. A correção prévia dos dados por hgi/hgm é uma forma de ponderar os dados de cada colheita por sua própria herdabilidade. Por outro lado, a correção Sm/Si dá peso 1,81 para a colheita com herdabilidade 7,6 % e peso 0,61 para a colheita com herdabilidade 35,4 %. Isto não é coerente. Assim, antes de se fazer a análise conjunta de colheitas para esse experimento, a correção hgi/hgm deve ser aplicada.

Avaliação em só local e em várias colheitas

Tomando-se os dados de todas as colheitas no experimento do Paraná, a análise conjunta de colheitas foi realizada segundo o modelo 55 do Selegen-Reml/Blup. Os principais resultados são apresentados a seguir.

(2) Análise Conjunta de Colheitas em um local (PR)

Tabela 54. Análise de deviance (ANADEV) referente à análise conjunta de colheitas no Paraná.

Efeito	Deviance	LRT (Qui-quadrado)	Comp.Var.	Coef. Determ.
Genótipos	349.4703^{+}	8.5919**	Vg= 0.1683**	h2g = 0.1815**
Permanente	425.5181^{+}	84.6397**	Vperm=0.2362**	c2perm =0.2547**
Genótipos X Colheitas	382.0803^{+}	41.2019**	Vgm=0.1699**	c2gm=0.1833**
Resíduo	-	-	Ve=0.3529	c2res = 0.3805
Modelo Completo	340.8784	-	-	c2total = 1.00

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente.
Repetibilidade de parcelas individuais = 0.4362
Correlação genética através das colheitas = 0.4976
+ Deviance do modelo ajustado sem os referidos efeitos

Verifica-se pela análise de deviance que os efeitos de genótipos, da interação genótipos x colheitas e de ambiente permanente e seus respectivos coeficientes de determinação foram significativos. Este elegante modelo de análise propiciou a estimação simultânea da herdabilidade individual (18,15 %), da repetibilidade individual (43,62 %) e da correlação genética através das colheitas (49,76 %), mesmo sob desbalanceamento. Os efeitos genotípicos explicaram 18,15 % e os

efeitos da interação com colheitas explicaram 18,33 % da variação total em colheitas de parcelas individuais. Essa interação é problemática para o melhorista, visto que a correlação do desempenho através das colheitas foi de moderada magnitude, com coincidência de aproximadamente 50 % dentre os melhores em todos os cortes. A magnitude (menor que aproximadamente 1) dessa correlação (0,4976) indica também que o caráter não é o mesmo de uma colheita para outra. Dessa forma, o modelo simplificado de repetibilidade sem a inclusão do efeito da interação genótipos x colheitas não é adequado.

A repetibilidade individual apresenta também moderada magnitude e informa que são necessários cinco cortes para se conseguir uma determinação de 80 % na avaliação do valor fenotípico permanente de uma parcela. A herdabilidade da média de genótipos através das seis colheitas (c) e dos três blocos (b) é dada por h2mg = Vg / (Vg + Vgm/c + Vperm / b + Ve / cb) = 0,5706. A raiz quadrada desse valor fornece a acurácia na seleção dos genótipos, a qual equivale a 75,54 %. Portanto, a experimentação realizada permitiu uma boa precisão, embora para testes de VCU a acurácia desejada deva ser em torno de 90 % ou mais. Os componentes de variância Vg, Vgm, Vperm e Ve referem-se às variâncias genotípica, da interação genótipos x medições, permanente e residual, respectivamente.

Esse modelo 55 de análise do Selegen-Reml/Blup fornece os valores genotípicos para a média das safras e também os valores genotípicos em cada corte, porém usando os dados de todas as safras simultaneamente. Esse último resultado é muito importante, sobretudo nesse caso em que a interação com cortes foi significativa e de natureza complexa, indicando que podem ser selecionados genótipos mais adaptados à estação seca e outros mais adaptados à estação das águas. Resultados para cada corte, usando simultaneamente a informação de todos os cortes não podem ser obtidos pelo método tradicional da análise de variância, mesmo quando os dados são completamente balanceados. Esta é mais uma vantagem da metodologia de modelos mistos (Reml/Blup). Assim, o método da análise de variância não tem lugar no melhoramento de forrageiras, pois sempre são realizadas várias colheitas nos experimentos.

O modelo 55 do Selegen-Reml/Blup fornece também a estabilidade (pelo método MHV-BLUP), a adaptabilidade (pelo método PRVG-BLUP) e a simultânea adaptabilidade e estabilidade dos valores genotípicos (MHPRVG-BLUP) através das safras.

Avaliação em vários locais e em uma só colheita

Tomando-se apenas os dados da primeira colheita nos três experimentos (Acre, Bahia e Paraná), a análise conjunta de locais foi realizada segundo o modelo 54 do Selegen-Reml/Blup. Os principais resultados são apresentados a seguir.

(3) Análise Conjunta de Locais (AC, BA, PR): Primeira Colheita

Tabela 55. Análise de deviance (ANADEV) referente à análise conjunta de locais na primeira colheita.

Efeito	Deviance	LRT (Qui-quadrado)	Comp.Var.	Coef. Determ.
Genótipos	258.7653+	0.1321 ns	Vg=0.0159 ns	h2g = 0.0172ns
Genótipos X Locais	260.7598+	2.1266 ns	Vint=0.0985 ns	c2int=0.1062 ns
Resíduo	-	-	Ve=0.8100	c2res = 0.8762
Modelo Completo	258.6332	-	-	c2total = 1.00

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente.
Correlação genética através dos locais = 0.1392
+ Deviance do modelo ajustado sem os referidos efeitos

Verifica-se pela análise de deviance que os efeitos de genótipos e da interação genótipos x locais e seus respectivos coeficientes de determinação não foram significativos. Isto mostra que a variabilidade fenotípica observada entre parcelas individuais através dos locais é basicamente de natureza ambiental, não havendo possibilidades de eficiente seleção de genótipos bons através de todos os locais e nem mesmo para locais específicos (visto que a interação com locais também foi não significativa), baseando-se apenas nos dados da primeira colheita.

Esse modelo 54 de análise do Selegen-Reml/Blup fornece os valores genotípicos para a média das locais e também os valores genotípicos em cada local, porém usando os dados de todas os locais simultaneamente. Esse último resultado é muito importante quando a interação com locais for significativa e de natureza complexa, indicando que podem ser selecionados genótipos mais adaptados à cada local. Resultados para cada local, usando simultaneamente a informação de todos

os locais, não podem ser obtidos pelo método tradicional da análise de variância, mesmo quando os dados são completamente balanceados. Esta é mais uma vantagem da metodologia de modelos mistos (Reml/Blup).

O modelo 54 do Selegen-Reml/Blup fornece também a estabilidade (pelo método MHV-BLUP), a adaptabilidade (pelo método PRVG-BLUP) e a simultânea adaptabilidade e estabilidade dos valores genotípicos (MHPRVG-BLUP) através dos locais. Um exemplo completo para esse caso em cana-de-açúcar foi apresentado no Capítulo 8.

Avaliação em vários locais e em várias colheitas

Tomando-se os dados de todas as colheitas nos três experimentos (Acre, Bahia e Paraná), a análise conjunta de locais e colheitas foi realizada segundo o modelo 155. Os principais resultados são apresentados a seguir.

(4) Análise Conjunta de Locais (AC, BA, PR) e Colheitas

Tabela 56. Análise de deviance (ANADEV) referente à análise conjunta global de locais e colheitas.

Efeito	Deviance	LRT(Qui-quadrado)	Comp.Var.	Coef. Determ.
Genótipos	1755.19^{+}	2.24 ^{ns}	0.0406^{ns}	h2g = 0.0366 ^{ns}
Genótipos	1760.17^{+}	7.22**	0.0732**	c2gl=0.0659**
X Locais				
Genótipos	1765.65^{+}	12.70**	0.0774**	c2gm=0.0697**
X Colheitas				
Genótipos	1805.59^{+}	52.64**	0.1775**	c2glm=0.1598**
X Locais				
X Colheitas				
Permanente	1836.22^{+}	83.27**	0.1429**	c2perm =0.1287**
Resíduo	-	-	0.5977	c2res=0.5388
Modelo Completo	1752.95	-	-	c2total=1.00

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente.
Repetibilidade de parcelas individuais = 0.2313
Correlação genética através das colheitas, válida para um novo local = 0.3443
Correlação genética através dos locais, válida para uma nova colheita = 0.3571
Correlação genética através das colheitas e dos locais = 0.1102
Correlação genética através das colheitas em um dado local = 0.5952
Correlação genética através dos locais em uma dada colheita = 0.6173
Correlação genética através das colheitas, considerando a média dos locais = 0.6160
Correlação genética através dos locais, considerando a média das colheitas = 0.5318
+ Deviance do modelo ajustado sem os referidos efeitos

Verifica-se pela análise de deviance que os efeitos da interação genótipos x locais, da interação genótipos x colheitas, da interação genótipos x locais x colheitas e de ambiente permanente e seus respectivos coeficientes de determinação foram significativos. Porém, os efeitos de genótipos não foram significativos, não havendo possibilidades de eficiente seleção de genótipos bons através de todos os locais e colheitas. Entretanto, para locais e/ou safras específicos existe a possibilidade de eficiente seleção (visto que todas as interações foram significativas).

Este elegante modelo de análise propiciou a estimação simultânea da herdabilidade individual (3,66 %), da repetibilidade individual (23,13 %), da correlação genética através das

colheitas (34,43 %), da correlação genética através dos locais (35,71 %), da correlação genética através das colheitas e locais (11,02 %), da correlação genética através das colheitas para um dado local (59,52 %), da correlação genética através dos locais para uma dada colheita (61,73 %), da correlação genética através das colheitas para a média de locais (61,60 %) e da correlação genética através dos locais para a média de colheitas (53,18 %), mesmo sob desbalanceamento. Os efeitos genotípicos livres de todas as interações explicaram apenas 3,66 % da variação total em colheitas de parcelas individuais nos vários locais. De todas as correlações apresentadas, as mais importantes para o melhorista são:

a) correlação genética através dos locais para a média de colheitas (53,18 %): apresentou magnitude moderada indicando uma coincidência de 53 % dos melhores nos vários locais, portanto, a seleção de genótipos específicos para cada local deve ser mais eficiente.

b) correlação genética através das colheitas para um dado local (59,52 %): apresentou magnitude moderada a alta indicando uma coincidência de 60 % dos melhores nas várias safras em cada local, portanto, a seleção de genótipos específicos para cada classe de colheita (secas e águas) pode ser mais eficiente. Genótipos selecionados para colheita da seca podem ser recomendados para regiões mais áridas e genótipos selecionados para colheita das águas podem ser recomendados para regiões mais úmidas.

c) correlação genética através das colheitas e locais (11,02 %): apresentou magnitude baixa indicando uma coincidência de apenas 11 % dos melhores através das várias safras e locais, portanto, a seleção de um genótipo bom para todos os locais e safras é improvável.

d) correlação genética através dos locais, válida para uma nova colheita (35,71 %): apresentou magnitude baixa indicando uma coincidência de 36 % dos melhores nos vários locais em uma nova safra futura (além das seis ou oito já realizadas) a ser avaliada, portanto, a seleção de genótipos específicos para cada local deve ser mais eficiente visando à produtividade em inúmeros cortes, que é a situação real sob pastejo nos sistemas de produção.

A repetibilidade individual apresenta baixa magnitude e informa que são necessários 13 cortes para se conseguir uma determinação de 80 % na avaliação do valor fenotípico permanente de uma parcela (resultado válido para um local qualquer). A herdabilidade da média de genótipos através das seis colheitas (c), dos três locais (l) e dos três blocos (b) é dada por h2mg = Vg / (Vg + Vgl/l + Vgm/c + Vglm/lc + Vperm / b + Ve / cbl) = 0,2772. A raiz quadrada desse valor fornece a acurácia na seleção dos genótipos, a qual equivale a 52,65 %. Portanto, a experimentação realizada não permite uma boa seleção para todos os locais e safras simultaneamente. Os componentes de variância Vgl e Vglm referem-se às variâncias da interação genótipos x locais e da interação genótipos x locais x medições, respectivamente.

Esse modelo 155 de análise do Selegen-Reml/Blup fornece os valores genotípicos para a média dos locais e também os valores genotípicos em cada local, porém usando os dados de todas os locais e safras simultaneamente. Esse último resultado é muito importante, sobretudo nesse caso em que a interação com locais foi significativa e de natureza complexa, indicando que podem ser selecionados genótipos mais adaptados a cada local. Resultados para cada local, usando simultaneamente a informação de todos os locais e cortes não podem ser obtidos pelo método tradicional da análise de variância, mesmo quando os dados são completamente balanceados. Esta é mais uma vantagem da metodologia de modelos mistos (Reml/Blup).

O modelo 155 do Selegen-Reml/Blup fornece também a estabilidade (pelo método MHV-BLUP), a adaptabilidade (pelo método PRVG-BLUP) e a simultânea adaptabilidade e estabilidade dos valores genotípicos (MHPRVG-BLUP) através dos locais.

Os resultados (valores genotípicos preditos, obtidos do arquivo com extensão .fam) referentes à seleção dos dez melhores acessos para todos os locais e safras simultaneamente são apresentados abaixo.

Ordem	Genótipo	VG	LIIC	LSIC
1	PM05	3.1988	3.1211	3.2766
2	PM07	3.0309	2.9532	3.1087
3	PM03	3.0276	2.9499	3.1054
4	PM08	3.0233	2.9456	3.1011
5	PM10	3.0131	2.9354	3.0909
6	PM04	3.0117	2.9339	3.0894
7	PM06	2.9970	2.9192	3.0747
8	PM14	2.9722	2.8944	3.0499
9	PM09	2.9543	2.8765	3.0321
10	PM23	2.9407	2.8629	3.0184

Verifica-se que o limite superior do intervalo de confiança do genótipo PM07, classificado em segundo lugar, é inferior ao limite inferior do intervalo de confiança do genótipo PM05, classificado em primeiro lugar. Assim, os dois genótipos diferem estatisticamente. Entretanto, isto não ocorre entre os nove indivíduos restantes. Isto é coerente com a não significância dos efeitos genotípicos através das colheitas e locais, relatada anteriormente. Como já comentado, a seleção deve ser realizada para cada local. Esta seleção é fornecida pelo Selegen mas não será apresentada aqui.

É importante relatar que os testes de comparação de médias não devem ser aplicados nessa situação. Primeiro, porque os efeitos genotípicos devem e foram tratados como aleatórios. Segundo, porque testes tradicionais como o de Tukey apresentam baixo poder de detecção de diferenças significativas quando o número de tratamentos é superior a cinco.

3.2 Avaliação de Indivíduos em Progênies de Meios Irmãos de *Brachiaria* em Várias Colheitas

A avaliação de progênies de meios irmãos é uma atividade comum nos programas de seleção recorrente em espécies perenes. Nesse caso, as unidades de recombinação são indivíduos e não famílias inteiras tal como ocorre no melhoramento de culturas anuais, em que são utilizadas sementes remanescentes para recombinação. Em espécies perenes, os próprios indivíduos avaliados são recombinados, podendo-se também incluir na recombinação alguns genitores. Assim, os modelos de análise devem ser ao nível de indivíduos e não ao nível de médias de famílias tal

como usado no melhoramento de plantas anuais. Este tipo de análise (nível de médias) apresenta deficiências, pois: não lida com o desbalanceamento dos dados, fato que sempre ocorre na experimentação de campo; não utiliza todos os efeitos do modelo estatístico estabelecido em nível de indivíduo; sob desbalanceamento não utiliza adequadamente o parentesco genético entre os indivíduos em avaliação; não considera que os próprios indivíduos avaliados serão recombinados e não os seus irmãos (sementes remanescentes), ou seja, não considera a coincidência entre unidade de seleção e unidade de recombinação.

O método ótimo de seleção é o BLUP individual, o qual utiliza todos os efeitos do modelo estatístico, contempla o desbalanceamento, utiliza o parentesco genético entre os indivíduos em avaliação, considera a coincidência entre unidade de seleção e unidade de recombinação. Na estimação dos parâmetros genéticos, o método da análise de variância não permite considerar o desbalanceamento. Somente o método da máxima verossimilhança residual (REML) permite uma análise eficiente nessa situação. Assim, o uso do método REML/BLUP é essencial nessa situação.

O presente tópico trata da avaliação genotípica de dez progênies de meios irmãos híbridas de *Brachiaria* em um experimento estabelecido no delineamento de blocos ao acaso com dez repetições e duas plantas por parcela. Foi avaliada a característica porcentagem de folhas em dois cortes ou colheitas de folhas. O experimento pode ser analisado por pelo menos dois modelos do Selegen-Reml/Blup: modelo 1 para análises individuais e modelo 62 para análise dos dois cortes em conjunto.

Resultados das Análises Individuais

Inicialmente cada corte foi analisado individualmente pelo modelo 1 do Selegen. Os resultados obtidos são relatados na seqüência.

Tabela 57. Estimativas de parâmetros genéticos e fenotípicos em cada corte em *Brachiaria.*

Estimativas de Parâmetros	Porcentagem de Folhas Corte 1	Porcentagem de Folhas Corte 2
Va	0.0201	0.0058
Vparc	0.0019	0.0029
Ve	0.0012	0.0243
Vf	0.0232	0.0330
h2a	0.8680	0.1744
h2aj	0.9438	0.1915
c2parc	0.0803	0.0889
h2mp	0.8341	0.4549
Acprog	0.9133	0.6744
h2ad	0.9265	0.1508
CVgi%	26.2622	16.2247
CVgp%	13.1311	8.1124
CVe%	18.5194	28.0838
CVr	0.7090	0.2889
PEV	0.0008	0.0008
SEP	0.0289	0.0280
Média Geral	**0.5398**	**0.4677**

As estimativas apresentadas na Tabela acima tem os seguintes significados:

Va: variância genética aditiva total na população do experimento.

Vparc: variância ambiental entre parcelas.

Ve : variância residual dentro de parcelas (ambiental + genética não aditiva).

Vf : variância fenotípica individual.

h2a : herdabilidade individual no sentido restrito, ou seja, dos efeitos aditivos.

h2aj: herdabilidade individual no sentido restrito, ajustada para os efeitos de parcela.

c2parc: coeficiente de determinação dos efeitos de parcela.

h2mp: herdabilidade da média de progênies, assumindo sobrevivência completa.

Acprog: acurácia da seleção de progênies, assumindo sobrevivência completa.

h2ad: herdabilidade aditiva dentro de parcela.

CVgi%: coeficiente de variação genética aditiva individual.

CVgp%: coeficiente de variação genotípica entre progênies.

CVe%: coeficiente de variação residual.

CVr = CVgp/Cve: coeficiente de variação relativa.

PEV: variância do erro de predição dos valores genotípicos de progênie, assumindo sobrevivência completa.

SEP: desvio padrão do valor genotípico predito de progênie, assumindo sobrevivência completa.

Média geral do experimento.

As seguintes conclusões podem ser emitidas com base nas estimativas dos parâmetros genéticos: (i) a média geral apresentou magnitude próxima para os dois caracteres; (ii) existe grande variabilidade genética para o caráter na população, conforme corroborado pela estimativa do CVgi% para ambos os caracteres; (ii) a precisão e qualidade do experimento foram altas para o primeiro corte, conforme demonstrado pelo CVr com estimativa de 0.71, valor este adequado para o caso de experimentos com dez repetições, no entanto, para o segundo corte a precisão foi mais baixa; (iii) a acurácia seletiva propiciada foi elevada (91 %) para o primeiro corte e moderada (67 %) para o segundo corte; (iv) o caráter apresenta controle genético alto no primeiro corte (86,8 %) e baixo (17,4 %) no segundo corte, conforme estimativa da herdabilidade individual; (v) a variação entre parcelas dentro de blocos explicou cerca de 8 % a 9 % da variabilidade total dentro de bloco para os dois caracteres, conforme revelado pelo c2parc, indicando que existe alguma heterogeneidade ambiental dentro de bloco, assim, blocos menores poderiam ser usados; (vi) o desvio padrão do valor genético predito (SEP) apresentou baixa magnitude (menos que 6 % da média, com coeficiente de variação do valor genético predito entre 5 % e 6 %), indicando boa precisão experimental e alta acurácia seletiva entre progênies. O parâmetro coeficiente de variação do valor genético predito é

muito mais informativo do que o coeficiente de variação experimental ou residual que, no caso, equivaleu a 18,5 % e 28,1 %. Isto porque o primeiro considera simultaneamente a variação residual, a variação genética e o número de repetições. O CVe% considera apenas a variação residual.

As herdabilidades entre (h2mp) e dentro (h2ad) de progênies equivaleram a 86,4 % e 92,6 % para o primeiro corte e 45,5 % e 15,1 % para o segundo corte. Esses (h2mp e h2ad) são os pesos dados à média corrigida de família e desvio corrigido do indivíduo dentro de família, respectivamente, para obtenção da predição BLUP dos valores genéticos individuais.

As estimativas da herdabilidade individual diferiram bastante para os dois cortes, em função de heterogeneidade de variância genética e residual. Assim, para realização da análise conjunta dos dois cortes, recomenda-se usar a transformação hi/hm.

Avaliação em várias colheitas

No presente caso, o modelo de análise conjunta envolvendo dados de diferentes cortes também precisa ser definido em nível de indivíduo. E nesse caso deve ser incorporado um vetor de efeitos aleatórios de ambiente permanente visando contemplar o fato de que as medidas repetidas (cortes sucessivos) no mesmo indivíduo são correlacionadas. Tal modelo permitirá estimar simultaneamente a herdabilidade e a repetibilidade dos caracteres, além de propiciar estimativas dos valores genéticos em nível individual, livres de todos os efeitos ambientais de parcelas, blocos e épocas de colheita.

O modelo completo de repetibilidade ou de simetria composta (CS) associado ao delineamento experimental de blocos ao acaso com várias plantas por parcela é dado por $Y_{ijkl} = \mu + g_i + b_j + m_k + gb_{ij} + gm_{ik} + bm_{jk} + gbm_{ijk} + e_{ijkl}$. Considerando os efeitos ambientais de blocos (b), medições (m) e interação blocos x medições como fixos, os mesmos podem ser ajustados somados a média geral, em um único vetor de efeitos fixos (β) dado pela combinação bloco-medição. Assim, o modelo linear misto resultante equivale a $Y_{ijkl} = \beta_{jk} + g_i + gm_{ik} + gb_{ij} + gbm_{ijk} + e_{ijkl}$. Esse modelo foi implementado como modelo 116 no *software* Selegen-Reml/Blup. Tal modelo pode ser usado como modelo inicial para testar a significância dos vários efeitos, via análise de deviance, conforme apresentado no Capítulo 3. O

modelo 116 não permite a seleção de indivíduos mas se o número de colheitas é o mesmo para todos os indivíduos, tal seleção pode ser realizada com eficiência máxima pelo modelo 1, entrando com os dados ao nível de médias por indivíduo. Nesse caso, recomenda-se o uso do modelo 116 para a estimação de parâmetros genéticos e o modelo 1 para a seleção de indivíduos.

Desdobrando este modelo em termos de efeitos permanentes (p) e temporários (t), tem-se $y = \beta + g_p + g_t + p_p + p_t + e_p + e_t$, em que:

$g_i = g_p$: efeito de genótipo, permanente através das colheitas.

$gm_{ik} = g_t$: efeito de genótipo, temporário em cada colheita.

$gb_{ij} = p_p$: efeito de parcela, permanente através das colheitas.

$gbm_{ijk} = p_t$: efeito de parcela, temporário em cada colheita.

$e_{ijk} = e_p + e_t$: efeito permanente + temporário de indivíduo dentro de parcela.

Em termos de variâncias destes efeitos, têm-se:

$\sigma^2_{gp} = \sigma^2_g$: variância genotípica ou covariância dos efeitos genotípicos através das colheitas; é a covariância genotípica através das colheitas em um modelo multivariado.

$\sigma^2_{gt} = \sigma^2_{gm}$: variância da interação genótipos x medições.

$\sigma^2_{pp} = \sigma^2_{gb}$: variância dos efeitos permanentes de parcela ou covariância dos efeitos de parcela através das colheitas em um modelo multivariado.

$\sigma^2_{pt} = \sigma^2_{gbm}$: variância dos efeitos temporários de parcela ou da interação parcelas x medições.

σ^2_{ep}: variância permanente de indivíduo dentro de parcela ou covariância dos efeitos de indivíduos dentro de parcela através das colheitas em um modelo multivariado.

σ^2_{et}: variância temporária de indivíduo dentro de parcela.

Verifica-se que tal modelo é bastante próximo ao modelo multivariado, desde que haja homogeneidade de variâncias. Assumindo que a interação dos ambientes da parcelas x medições (gbm) é desprezível e/ou pode ser reunido ao erro temporário, o modelo simplifica-se para $y = \beta + g_p + g_t + p_p + e_p + e_t = \beta + g + gm + gb + e_p + e_t$, o qual é denominado modelo de repetibilidade + interação genótipos x medições. Esse modelo foi implementado como modelo 62 no Selegen-Reml/Blup. Assumindo adicionalmente que a correlação genotípica através das medições aproxima 1 (interação g x m não significativa), o modelo se reduz a $y = \beta + g_p + p_p + e_p + e_t = \beta + g + gb + e_p + e_t$, o qual é denominado modelo simplificado de repetibilidade, que se encontra implementado nos modelos 8 e 67 do Selegen. A adequação desses modelos pode ser avaliada rodando-se o modelo 116, declarando o coeficiente c2gm como zero e realizando o teste da razão de verossimilhança, conforme descrito no Capítulo 3.

Tomando-se os dados das duas colheitas no experimento, a análise conjunta de colheitas foi realizada segundo o modelo 62 do Selegen. Os principais resultados são apresentados a seguir.

Tabela 58. Análise de deviance (ANADEV) referente à análise conjunta de colheitas em *Brachiaria.*

Efeito	Deviance	LRT (Qui-quadrado)	Comp.Var.	Coef. Determ.
Progênies	-944.8563$^+$	0.00^{ns}	Vg= 0.00001^{ns}	h2g = 0.0003^{ns}
Parcela	936.5771$^+$	8.28**	Vparc=0.00656**	c2parc=0.2133**
Progênies	921.4699$^+$	23.39**	Vgm=0.00336**	c2gm= 0.1092**
X Colheitas				
Permanente	944.7366$^+$	0.12^{ns}	Vperm= 0.00057^{ns}	c2perm = 0.0185^{ns}
Resíduo	-	-	Ve=0.02027	c2res = 0.6584
Modelo Completo	-944.8563	-	-	c2total = 1.00

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente
Repetibilidade de parcelas individuais = 0.2322
Correlação genética através das colheitas = 0.0031
+ Deviance do modelo ajustado sem os referidos efeitos.

Os significados das estimativas apresentadas na Tabela acima são:

Vg: variância genotípica entre progênies; equivale a (1/4) da variação genética aditiva.

Vparc: variância ambiental entre parcelas.

Vgm: variância da interação progênies x medições.

Vperm: variância dos efeitos permanentes.

Ve: variância residual temporária.

Vf: variância fenotípica individual.

h2g: herdabilidade entre progênies em nível individual.

r: repetibilidade individual.

c2parc: coeficiente de determinação dos efeitos de parcela.

c2gm: coeficiente de determinação dos efeitos da interação genótipos x medições.

c2perm: coeficiente de determinação dos efeitos permanentes.

rgmed: correlação genotípica através das medições.

Média geral do experimento.

Verifica-se pela análise de deviance que os efeitos de progênies livres da interação com cortes e os efeitos de ambiente permanente do indivíduo não foram significativos. Por outro lado, os efeitos da interação genótipos x colheitas e de parcela e seus respectivos coeficientes de determinação foram significativos. Este elegante modelo de análise propiciou a estimação simultânea da herdabilidade individual (0,03 %), da repetibilidade individual (23,22 %) e da correlação genética através das colheitas (0,03 %), mesmo sob desbalanceamento. Os efeitos da interação de progênies com colheitas explicaram 10,92 % da variação total em colheitas de plantas individuais. Essa interação é problemática para o melhorista, visto que a correlação do desempenho através das colheitas foi praticamente nula, com ausência de coincidência dentre os melhores em

todos os cortes. A magnitude (menor que 1) dessa correlação indica também que o caráter não é o mesmo de uma colheita para outra. Dessa forma, o modelo simplificado de repetibilidade sem a inclusão do efeito da interação genótipos x colheitas não é adequado. A seleção deve ser realizada especificamente para cada corte, ou seja, uma seleção para o corte das águas e outra para o corte das secas.

A repetibilidade individual apresenta também baixa magnitude e informa que são necessários 15 cortes para se conseguir uma determinação de 82 % na avaliação do valor fenotípico permanente de um indivíduo. A herdabilidade para a seleção de indivíduos em cada corte nessa análise conjunta é dada por h2a = 4 x (h2g + c2gm) e equivale a 43,68 %.

Esse modelo 62 de análise do Selegen-Reml/Blup fornece os valores genotípicos das progênies para a média das safras e também os valores genotípicos em cada corte, porém usando os dados de todas as safras simultaneamente. Esse último resultado é muito importante, sobretudo nesse caso em que a interação com cortes foi significativa e de natureza complexa, indicando que podem ser selecionados genótipos mais adaptados à estação seca e outros mais adaptados à estação das águas. Resultados para cada corte, usando simultaneamente a informação de todos os cortes não podem ser obtidos pelo método tradicional da análise de variância mesmo quando os dados são completamente balanceados. Esta é mais uma vantagem da metodologia de modelos mistos (Reml/Blup). Assim, o método da análise de variância não tem lugar no melhoramento de forrageiras, pois sempre são realizadas várias colheitas nos experimentos.

O modelo 62 do Selegen-Reml/Blup fornece também a estabilidade (pelo método MHV-BLUP), a adaptabilidade (pelo método PRVG-BLUP) e a simultânea adaptabilidade e estabilidade dos valores genotípicos (MHPRVG-BLUP) através das safras.

Os valores genotípicos de progênies em cada corte são apresentados a seguir.

Tabela 20. Valores genotípicos preditos de progênies de *Brachiaria* no primeiro corte.

Ordem	**Progênie**	**Valor Genotípico Corte 1**
1	BS15	0.6325
2	BS9	0.6073
3	BS13	0.5419
4	BS14	0.5121
5	BS17	0.4966
6	BS11	0.4884
7	BS18	0.4833
8	BS7	0.4829
9	BS6	0.4698
10	BS3	0.4661
	Média Geral	**0.5181**
Ordem	**Progênie**	**Valor Genotípico Corte 2**
1	BS3	0.5197
2	BS6	0.511
3	BS18	0.491
4	BS15	0.4906
5	BS7	0.4837
6	BS13	0.4618
7	BS11	0.4421
8	BS17	0.4384
9	BS9	0.4346
10	BS14	0.4037
	Média Geral	0.4677

Verifica-se que a melhor progênie no corte 2 é a pior no corte 1, conforme esperado de acordo com a grande interação progênies x cortes verificada. A melhor progênie no corte 1 apresentou 63,25 % de folhas. No corte 2, a mesma progênie apresentou 49,1 % de folhas.

O modelo 62 do Selegen fornece também a seleção dos indivíduos para cada corte, considerando simultaneamente os dois cortes. A seguir são apresentados os cinco melhores indivíduos para cada corte.

Tabela 60. Valores genéticos aditivos individuais de indivíduos de *Brachiaria* em cada corte.

Ordem	Bloco	Família	Árvore	Fenótipo	Valor Aditivo Corte 1
1	8	BS9	2	0.9864	0.7221
2	4	BS15	2	0.913	0.7053
3	6	BS15	1	0.8478	0.7037
4	7	BS15	2	0.9481	0.7033
5	10	BS15	1	0.793	0.6993
Ordem	**Bloco**	**Família**	**Árvore**	**Fenótipo**	**Valor Aditivo Corte 2**
1	7	BS7	1	1	0.696
2	1	BS3	1	0.88	0.6839
3	7	BS6	1	0.93	0.6723
4	3	BS3	1	0.81	0.6701
5	10	BS6	2	0.75	0.6698

Verifica-se a diferença entre os valores fenotípicos e os valores genéticos aditivos, exatamente devido à presença de efeitos ambientais nos fenótipos. A seleção dos melhores indivíduos no experimento poderá elevar a porcentagem de folhas de cerca de 50 % para cerca de 70 %, revelando o grande potencial do melhoramento em aumentar a produtividade forrageira.

Verifica-se também que no corte 1 a maioria dos melhores indivíduos são da progênie BS15. Assim, para recombinação dos melhores indivíduos da população experimental, deve haver alguma restrição no número máximo de indivíduos contribuídos por família, como forma de evitar o comprometimento do efetivo populacional e a ocorrência de endogamia na população de melhoramento. Essa otimização da seleção pode ser realizada empregando-se o modelo 106 do Selegen-Reml/Blup.

3.3 Avaliação Multivariada Envolvendo Acessos de *Stylosanthes*: Estrutura de Correlações, Divergência Genética e Índice de Seleção

O presente tópico trata da avaliação genotípica de 35 acessos de *Stylosanthes guianensis* em um experimento estabelecido no delineamento de blocos ao acaso com quatro repetições e três

plantas por parcela. Foram avaliadas onze características morfológicas. O experimento é desbalanceado devido à perda de algumas plantas. O experimento pode ser analisado por pelo menos dois modelos do Selegen-Reml/Blup: modelo 2 se somente há o interesse na avaliação dos acessos ou modelo 24 se há também o interesse em avaliar indivíduos dentro dos acessos. No presente exemplo, foi utilizado o modelo 2, cuja seqüência de colunas equivale à mesma descrita para o modelo 24.

Resultados das Análises Individuais

Inicialmente cada variável foi analisada individualmente. Os resultados obtidos são relatados na seqüência, tomando como exemplo a terceira variável, denominada número de ramos por planta.

Vg = 136,8149: variância genotípica.

Vparc = 53,1919: variância ambiental entre parcelas.

Ve = 230,2183: variância residual.

Vf = 420,2252: variância fenotípica individual.

h2g = 0,3255: herdabilidade individual no sentido amplo, ou seja, dos efeitos genotípicos totais.

h2aj = 0,3727: herdabilidade individual no sentido amplo, ajustada para os efeitos de parcela.

c2parc = 0,1266: coeficiente de determinação dos efeitos de parcela.

h2mc = 0,8081: herdabilidade da média de genótipo, assumindo ausência de perda de parcelas.

Acclon = 0,8989: acurácia da seleção de genótipos, assumindo ausência de perda de parcelas.

CVgi% = 22,0897: coeficiente de variação genotípica.

CVe% = 21,5269: coeficiente de variação residual.

CVr = CVg/CVe = 1,0261: coeficiente de variação relativa.

PEV = 26,2504: variância do erro de predição dos valores genotípicos, assumindo sobrevivência completa.

SEP = 5,1235: desvio padrão do valor genotípico predito, assumindo sobrevivência completa.

Média geral do experimento = 52,9511.

As seguintes conclusões podem ser emitidas com base nas estimativas dos parâmetros genéticos: (i) existe grande variabilidade genética para o caráter na população, conforme corroborado pela estimativa do CVgi%; (ii) a precisão e a qualidade do experimento foram altas, conforme demonstrado pelo CVr com estimativa maior que 1, valor este adequado para o caso de experimentos com quatro repetições, conforme mostrado no Capítulo 3; (iii) a acurácia seletiva propiciada foi elevada (90 %), adequada para ensaios de VCU, conforme o Capítulo 3; (iv) o caráter apresenta controle genético moderado (maior que 30 %) a alto (maior que 40 %), conforme estimativa da herdabilidade individual; (v) a variação entre parcelas dentro de blocos explicou 12,66 % da variabilidade total dentro de bloco, conforme revelado pelo c2parc, indicando que existe considerável heterogeneidade ambiental dentro de bloco; (vi) o desvio padrão do valor genético predito (SEP) apresentou baixa magnitude (menos que 10 % da média, com coeficiente de variação do valor genético predito igual 9,67 %), indicando boa precisão experimental e alta acurácia seletiva. O parâmetro coeficiente de variação do valor genético predito é muito mais informativo do que o coeficiente de variação experimental ou residual, que no caso, equivaleu a 21,53 %. Isto porque o primeiro considera simultaneamente a variação residual, a variação genotípica e o número de repetições. O CVe% considera apenas a variação residual.

Correlações Genéticas Entre Caracteres

Após a realização das análises individuais para os onze caracteres, as correlações genotípicas entre os caracteres foram obtidas pelo modelo 102 do Selegen-Reml/Blup. A matriz de correlações é apresentada a seguir.

Tabela 61. Matriz de correlação entre caracteres de *Stylosanthes.*

Variável	1	2	3	4	5	6	7	8	9	10	11
1	1	-0.4214	0.8898	0.0096	-0.1119	0.0392	-0.2953	0.0346	-0.2659	0.0841	-0.0322
2	-0.4214	1	-0.3581	-0.3186	0.5671	-0.1598	0.4587	-0.1695	0.3867	0.0477	0.1161
3	0.8898	-0.3581	1	-0.0113	0.1193	0.0414	-0.2775	0.0486	-0.2672	0.0367	0.0158
4	0.0096	-0.3186	-0.0113	1	-0.3876	-0.0706	-0.2202	-0.2027	-0.1937	-0.5234	-0.0565
5	-0.1119	0.5671	0.1193	-0.3876	1	0.1027	0.1789	0.1445	0.1458	0.1698	0.1977
6	0.0392	-0.1598	0.0414	-0.0706	0.1027	1	0.102	0.9546	0.2557	0.6378	0.1996
7	-0.2953	0.4587	-0.2775	-0.2202	0.1789	0.102	1	0.045	0.9694	0.4946	0.3527
8	0.0346	-0.1695	0.0486	-0.2027	0.1445	0.9546	0.045	1	0.2083	0.7116	0.121
9	-0.2659	0.3867	-0.2672	-0.1937	0.1458	0.2557	0.9694	0.2083	1	0.5842	0.3307
10	0.0841	0.0477	0.0367	-0.5234	0.1698	0.6378	0.4946	0.7116	0.5842	1	0.2114
11	-0.0322	0.1161	0.0158	-0.0565	0.1977	0.1996	0.3527	0.121	0.3307	0.2114	1

Verifica-se que todas as correlações encontram-se dentro do espaço paramétrico (-1 a 1). As correlações de maiores magnitudes foram obtidas entre os caracteres 1 e 3 (0,89), 6 e 8 (0,95) e 7 e 9 (0,97). Correlações maiores que 0,80 praticamente não contribuem para explicar a variabilidade total entre genótipos, além de causarem problemas de multicolinearidade em procedimentos de estimação. Assim, recomenda-se descartar um dos caracteres de cada par citado, antes da realização das análises multivariadas nos estudos de divergência genética.

Análises Multivariadas para Estudos de Divergência Genética e Agrupamento de Genótipos

De posse da matriz de correlações genotípicas e da matriz de valores genotípicos preditos foram efetuadas a análise de componentes principais genéticos (modelo 103 do *software* Selegen-Reml/Blup) e a análise de agrupamento pelo método Tocher (modelo 104 do software Selegen-Reml/Blup), usando três medidas de distância genética: distância euclidiana média genética, distância euclidiana quadrada genética e distância estatística ou de Mahalanobis genética.

Para a realização das análises, foram eliminadas as variáveis 1, 8 e 9, as quais apresentaram altas correlações genéticas com as variáveis 3, 6 e 7, respectivamente. Essas últimas foram mantidas por exibirem maior herdabilidade e variação genética e, portanto, serem mais importantes na diversidade em nível genético.

Análise de Componentes Principais

A análise de componentes principais revelou os seguintes resultados.

Tabela 62. Resultados da análise de componentes principais em *Stylosanthes.*

Componente	Autovalores	Proporção Explicada	Proporção Acumulada
1	2.5871	0.3234	0.3234
2	1.6525	0.2066	0.5299
3	1.2249	0.1531	0.6831
4	0.9844	0.1230	**0.8061**
5	0.7018	0.0877	0.8938
6	0.5256	0.0657	0.9595
7	0.2126	0.0266	0.9861
8	0.1112	0.0139	1.0000

Os autovalores informam sobre a variância associada a cada componente principal. Os quatro primeiros componentes principais explicaram 80,61 % da variabilidade total explicada pelos oito caracteres. Assim, somente essas quatro novas variáveis (os componentes principais) podem ser usados para inferências práticas. Isto porque quando os componentes, de maneira acumulada, explicam mais que 80 % da variação total, toda a informação disponível pode ser resumida nesses componentes. Os escores de cada genótipo, associados a esses componentes principais, podem ser dispersos graficamente visando ao agrupamento de genótipos em grupos divergentes. Essa dispersão considera os componentes principais aos pares e permite verificar a consistência do agrupamento nas várias dispersões. Os escores dos genótipos nos dois primeiros componentes principais a serem usados para a dispersão em dois eixos são apresentados a seguir, para dez genótipos.

Tabela 63. Escores associados aos componentes principais em *Stylosanthes.*

Genótipos	Escores para o Componente 1	Escores para o Componente 2
4311	1.5212	2.5446
2768	-1.1425	-1.3297
4310	-1.4677	-0.8715
444	1.8579	1.0775
1371	-1.0965	-0.7579
4234	0.6954	1.5981
4194	-0.6711	-1.2182
4199	2.8329	1.7065
135	-0.0826	-0.2775
4306	-0.0565	1.2862

Na seqüência são apresentados os autovetores com os coeficientes de ponderação das variáveis em cada componente principal. Esses coeficientes permitem a obtenção dos escores dos componentes principais para cada genótipo ou acesso, por meio da multiplicação dos mesmos pelo valor observado de cada variável e posterior soma para todas as variáveis em um genótipo.

Tabela 64. Autovetores com os coeficientes de ponderação das variáveis nos componentes principais.

Var	Autovetor 1	Autovetor 2	Autovetor 3	Autovetor 4	Autovetor 5	Autovetor 6	Autovetor 7	Autovetor 8
2	0.3770	-0.5400	-0.0652	-0.0022	-0.2071	0.1461	**-0.6915**	-0.1401
3	-0.1123	0.3726	-0.6144	-0.3124	0.2941	0.4324	-0.3062	0.0797
4	-0.3994	-0.0063	0.3656	-0.4164	-0.3965	0.4923	0.0502	-0.3624
5	0.3736	-0.1954	-0.5230	-0.2075	-0.4280	0.0707	0.5607	-0.0557
6	0.2481	0.5644	0.1481	0.0521	-0.6012	-0.0044	-0.2270	0.4270
7	0.4461	-0.1153	0.3856	-0.0887	0.3520	**0.5486**	0.2280	0.3928
10	0.4562	0.4341	0.0811	0.2636	0.1253	0.1172	0.0314	**-0.7049**
11	0.2783	0.1053	0.1840	-0.7784	0.1709	-0.4792	-0.0820	-0.0778

Os maiores coeficientes das variáveis nos últimos componentes principais (aqueles com autovalores menores que 0.70) podem ser usados para o descarte de variáveis redundantes que pouco contribuem para a discriminação dos genótipos. No presente caso, deve-se considerar os últimos três autovetores e as variáveis passíveis de descarte são 10 (conforme revelado no último autovetor), 2 (conforme revelado no penúltimo autovetor) e 7 (conforme revelado no antepenúltimo autovetor). Esse descarte pode se basear também nas correlações entre cada variável e os componentes principais, conforme apresentado a seguir.

Tabela 65. Correlações entre cada variável e os componentes principais em *Stylosanthes.*

Var	CP 1	CP 2	CP 3	CP 4	CP 5	CP 6	CP 7	CP 8
2	0.6119	-0.6895	-0.0702	-0.0003	-0.1755	0.1034	**-0.3196**	-0.0402
3	-0.1812	0.4799	-0.6847	-0.3005	0.2424	0.3135	-0.1407	0.0311
4	-0.6430	-0.0094	0.3981	-0.4166	-0.3332	0.3584	0.0174	-0.1208
5	0.6044	-0.2484	-0.5805	-0.1914	-0.3602	0.0569	0.2565	-0.0230
6	0.3957	0.7285	0.1713	0.0524	-0.4997	-0.0062	-0.0972	0.1462
7	0.7156	-0.1494	0.4271	-0.0949	0.2974	**0.3964**	0.1078	0.1298
10	0.7296	0.5627	0.0949	0.2595	0.1059	0.0876	0.0051	**-0.2364**
11	0.4484	0.1362	0.1853	-0.7774	0.1390	-0.3469	-0.0360	-0.0265

Verifica-se que foram obtidas as mesmas conclusões com base nos coeficientes das variáveis nos autovetores. As variáveis de maiores correlações com os três últimos componentes principais são 10, 2 e 7. De maneira consistente entre as duas abordagens, essas variáveis podem ser descartadas em estudos futuros. É importante relatar que a abordagem das correlações inclui também efeitos indiretos em outras variáveis. Sendo assim, a primeira abordagem é mais indicada.

Os caracteres não passíveis de descarte são somente aqueles com alta variabilidade entre os materiais genéticos e com correlações baixas ou nulas com as outras variáveis.

Análise de Agrupamento

A análise de agrupamento pelo método Tocher (modelo 104), usando três medidas de distância genética: distância euclidiana média genética, distância euclidiana quadrada genética e distância estatística ou de Mahalanobis genética revelou resultados idênticos pelas duas primeiras medidas mencionadas. O resultado desse agrupamento é apresentado a seguir.

Tabela 66. Agrupamento pelo método de Tocher aplicado sobre distâncias euclidianas genéticas.

Grupo	Genótipos
1	Maioria
2	4199 e 1230
3	4310 e 4285
4	444

Quatro grupos foram formados, sendo que a maioria dos genótipos foi alocada no grupo 1. Genótipos de grupos distintos podem ser cruzados visando à obtenção de maior variabilidade genética na descendência ou para obtenção de uma possível heterose na descendência, para caracteres que exibem dominância alélica. Para esses objetivos, preferencialmente deve-se cruzar o indivíduo do grupo 4 com indivíduos do grupo 1.

Índices de Seleção para Vários Caracteres

De maneira genérica, o caráter de maior importância no melhoramento de forrageiras é a produção total de proteínas digestíveis na rebrota (PTPDR). Esse caráter é dado pelo produto dos caracteres produção de massa foliar (PMF), teor de proteínas (%P), porcentagem de digestibilidade (%D) e porcentagem de rebrota (%R), ou seja, PTPDR = PMF * %P * %D * %R. Tal caráter é fortemente correlacionado com o ganho de peso pelo animal. A %R deve ser incluída como forma de considerar indiretamente o caráter capacidade de suporte da pastagem. Em alguns casos, a própria PMF em várias safras já contempla a %R. Para a composição de índices de seleção a adequada

determinação dos pesos econômicos dos caracteres pode ser baseada nas correlações genéticas envolvendo os vários caracteres e a PTPDR, pela expressão peso caráter i = (correlação (i, PTPDR)) / (soma das correlações entre cada caráter e a PTPDR).

O modelo 101 do Selegen permite a construção de índices de seleção dos tipos multiplicativo, aditivo com pesos econômicos e índice com base em *ranking* médio. Esses tipos de índice foram aplicados considerando os caracteres número de ramos (NR), comprimento da haste (CH) e resistência à antracnose (RA, variando de 0 a 1, em que 1 indica resistência máxima). O produto desses três caracteres fornece a produtividade total de forragem sadia de uma planta e esse caráter composto equivale exatamente ao índice multiplicativo. Os resultados desse índice para os dez melhores genótipos são apresentados a seguir.

Tabela 67. Índice multiplicativo para *Stylosanthes.*

Ordem	Genótipo	Índice	Valor Genético CH	Valor Genético NR	Valor Genético RA
1	1109	458.41	82.267	6.547	0.8511
2	1585	416.88	77.877	6.601	0.8109
3	4171	360.85	61.911	7.369	0.7909
4	4310	356.15	36.330	10.303	0.9514
5	4233	350.41	64.191	7.470	0.7307
6	4227	346.69	61.424	7.136	0.7909
7	2769	346.15	58.418	6.802	0.8711
8	2757	342.33	59.234	6.954	0.8310
9	2736	340.52	57.879	6.601	0.8912
10	4144	339.09	56.362	6.906	0.8711
Média	-	299.13	49.925	7.013	0.8554

Verifica-se que os melhores genótipos apresentam certo equilíbrio nos três caracteres mas, ao mesmo tempo, esse índice permite que a superioridade em um caráter corrija a deficiência em outro. Parece que CH é mais importante, sendo que apenas um dos dez melhores, apresentou CH menor que a média, sendo nesse caso, compensado pela maior % de resistência à antracnose (95,14%). A seguir é apresentado o índice dado pelo *ranking* médio.

Tabela 68. Índice de *rank* médio para *Stylosanthes.*

Ordem	Genótipo	Ordem Média	Ordem CH	Ordem NR	Ordem RA
1	4310	12.00	33	1	2
2	4285	12.67	27	5	6
3	1371	13.33	28	3	9
4	4316	13.33	25	10	5
5	4171	14.00	4	7	31
6	2769	14.00	8	20	14
7	4306	14.33	15	9	19
8	2464	14.33	29	2	12
9	4233	14.33	3	6	34
10	4144	14.33	10	16	17

Verifica-se que cinco dos dez melhores são coincidentes pelos dois métodos (índice multiplicativo e índice de rank). Esse método privilegia genótipos equilibrados nos vários caracteres. A seguir são apresentados resultados do índice aditivo com pesos iguais para todos os caracteres.

Tabela 69. Índice aditivo com pesos iguais para *Stylosanthes.*

Ordem	Genótipo	Índice
1	4310	9.8895
2	4285	9.0555
3	1371	9.0208
4	2464	9.0078
5	1109	8.8913
6	1352	8.7760
7	4316	8.7321
8	4311	8.6337
9	2736	8.6205
10	2203	8.6106
Média	-	8.4101

Verifica-se que os três melhores genótipos foram idênticos pelo índice do *ranking* médio e pelo índice aditivo com pesos econômicos iguais. Isto revela que tais índices são similares.

A seguir são apresentados resultados obtidos pelo índice aditivo privilegiando o caráter RA.

Tabela 70. Índice aditivo privilegiando o caráter RA em *Stylosanthes.*

Ordem	Genótipo	Índice	RA
1	4310	12.3678	0.9514
2	1352	11.8277	0.9715
3	4285	11.5937	0.9113
4	2203	11.4581	0.9313
5	1371	11.4380	0.8912
6	2464	11.4302	0.8912
7	4316	11.3978	0.9113
8	4311	11.3160	0.9080
9	2768	11.3145	0.9113
10	664	11.2799	0.9113
Média	-	10.8283	0.8554

No índice acima foram dados os pesos 0.66 para RA e 0.33 para NR e CH. Verifica-se que os melhores genótipos apresentam resistência a antracnose superior a 89 %. O grupo dos dez melhores pelo índice multiplicativo admitiu genótipos com resistência acima de 79 %. A seguir são apresentados resultados obtidos pelo índice aditivo privilegiando o caráter NR.

Tabela 71. Índice aditivo privilegiando o caráter NR em *Stylosanthes.*

Ordem	Genótipo	Índice	NR
1	4310	10.6805	10.3036
2	1371	9.3217	8.4737
3	2464	9.3210	8.4895
4	4285	9.2049	8.1708
5	4316	8.5688	7.2032
6	4234	8.5564	8.2062
7	4332	8.5059	7.2758
8	4306	8.4941	7.2701
9	4171	8.4668	7.3694
10	1109	8.3668	6.547
Média	-	8.2873	

Verifica-se uma coincidência de 50 % entre os índices privilegiando RA e NR, indicando que não há muito prejuízo em priorizar um dos caracteres quando necessário for. Isto é reflexo da correlação genética praticamente nula entre as duas variáveis. Pode-se priorizar um dos caracteres, mas ambos devem ser considerados simultaneamente na seleção, visando encontrar recombinantes favoráveis em ambos os caracteres. De maneira alguma a seleção poderá basear-se em apenas um dos caracteres.

4 ESPÉCIES ANUAIS GRANÍFERAS E OLERÍCOLAS

Neste tópico são descritas algumas aplicações em Agricultura e Olericultura. São aplicáveis às gramíneas (milho, arroz, trigo, cevada, aveia, sorgo), leguminosas (feijão, soja), olerícolas (feijão de vagem, tomate, batata, mandioca, cenoura, cebola, alho, cucurbitáceas, brócolis, espinafre, agrião, etc) e também ao algodão. Como exemplo, serão considerados resultados do programa de melhoramento genético do feijoeiro do Instituto Agronômico de Campinas (IAC), conforme publicado por Carbonell et al. (2007).

4.1 Estimativas de Parâmetros Genéticos em Fejoeiro

Foram avaliadas 20 linhagens em 15 ambientes do Estado de São Paulo. Em cada ambiente foi estabelecido um experimento no delineamento de blocos ao acaso com quatro repetições. O modelo estatístico para análise dessa rede experimental considerando a tomada de uma observação por parcela é dado por:

y = Xb + Zg + Wc + e, em que:

y, b, g, ge, e = vetores de dados, de efeitos fixos (médias de blocos através dos locais), de efeitos genotípicos de linhagens (aleatório), de efeitos da interação genótipos x ambientes (aleatório) e de erros aleatórios, respectivamente.

X, Z e W = matrizes de incidência para b, g e ge, respectivamente.

Distribuições e estruturas de médias e variâncias:

$$E\begin{bmatrix} y \\ g \\ ge \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \\ 0 \end{bmatrix}; \qquad Var\begin{bmatrix} g \\ ge \\ e \end{bmatrix} = \begin{bmatrix} I\sigma_g^2 & 0 & 0 \\ 0 & I\sigma_{ge}^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix}$$

Equações de modelo misto:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + I\lambda_1 & Z'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{g}e \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$

em que:

$\lambda_1 = \frac{\sigma_e^2}{\sigma_g^2} = \frac{1 - h_g^2 - c_{ge}^2}{h_g^2}$; $\lambda_2 = \frac{\sigma_e^2}{\sigma_{ge}^2} = \frac{1 - h_g^2 - c_{ge}^2}{c_{ge}^2}$.

$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2 + \sigma_e^2}$ = herdabilidade individual no sentido amplo de parcelas individuais no bloco;

$c_{ge}^2 = \frac{\sigma_{ge}^2}{\sigma_g^2 + \sigma_{ge}^2 + \sigma_e^2}$: coeficiente de determinação dos efeitos da interação genótipos x ambientes;

σ_g^2 = variância genotípica entre linhagens;

σ_{ge}^2 = variância da interação genótipos x ambientes;

σ_e^2 = variância residual entre parcelas;

$r_{gloc} = \frac{\sigma_c^2}{\sigma_g^2 + \sigma_{ge}^2} = \frac{h_g^2}{h_g^2 + c_{ge}^2}$: correlação genotípica dos materiais genéticos através dos ambientes.

Estimadores de componentes de variância por REML via algoritmo EM:

$\hat{\sigma}_e^2 = [y'y - \hat{b}'\ X'y - \hat{g}'\ Z'y - \hat{c}'\ W'y] / [N - r(x)]$

$\hat{\sigma}_g^2 = [\hat{g}'\hat{g} + \hat{\sigma}_e^2\ tr\ C^{22})] / q$

$\hat{\sigma}_{ge}^2 = [g\hat{e}'g\hat{e} + \hat{\sigma}_e^2\ tr\ C^{33}] / s$

em que:

C^{22} e C^{33} advém de :

$$C^{-1} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}^{-1} = \begin{bmatrix} C^{11} & C^{12} & C^{13} \\ C^{21} & C^{22} & C^{23} \\ C^{31} & C^{32} & C^{33} \end{bmatrix}$$

C = matriz dos coeficientes das equações de modelo misto;

tr = operador traço matricial;

r(x) = posto da matriz X;

N,q,s = número total de dados, número de clones e número de combinações genótipos x ambientes, respectivamente.

Nesse modelo, os valores genotípicos preditos livres da interação considerando todos os locais são dados por u + g, em que u é a média de todos os locais. Para cada local j, os valores genotípicos são preditos por u_j + g + ge, em que u_j é a média do local j.

A seleção conjunta por produtividade, estabilidade e adaptabilidade dos materiais genéticos baseou-se na estatística denominada média harmônica da performance relativa dos valores genéticos (MHPRVG) preditos, conforme descrito por Resende (2004). A MHPRVG conduz a resultados semelhantes aos obtidos pelos métodos descritos por Lin e Binns (1988) e Annicchiarico (1992). Todas as análises foram realizadas por meio do *software* Selegen-Reml/Blup, empregando o modelo 54 (Resende, 2002).

Os resultados referentes aos componentes de variância e parâmetros genéticos para a análise conjunta dos 15 ambientes e análise conjunta de épocas e anos de plantio para os locais 1 e 2 são apresentados na Tabela 72, para o caráter produção de grãos.

Tabela 72. Estimativas dos coeficientes de herdabilidade individual no sentido amplo ($\hat{h}_g^2$), coeficiente de determinação dos efeitos da interação genótipos x ambientes ($\hat{c}_{ge}^2$), herdabilidade da média de linhagem ($\hat{h}_{ml}^2$), variância genotípica ($\hat{\sigma}_g^2$), variância da interação genótipos x ambientes ($\hat{\sigma}_{ge}^2$), variância residual entre parcelas ($\hat{\sigma}_e^2$), variância fenotípica individual ($\hat{\sigma}_f^2$), correlação genotípica através dos locais ($\hat{r}_{gloc}$), acurácias na seleção de linhagens ($\hat{r}_{g\hat{g}}$), média geral, coeficiente de variação genética (cv_g%) e coeficiente de variação experimental (cv_e%), para o caráter produção de grãos em feijoeiro, na análise de 15 ambientes no Estado de São Paulo e também análise conjunta de duas épocas e dois anos para os locais 1 e 2.

Estimativas	**Ambiente 1 (Capão Bonito)**	**Ambiente 2 (Masul)**	**Análise Conjunta (15 ambientes)**
	Produção	**Produção**	**Produção**
($\hat{h}_g^2$)	0.1643± 0.0641	0.3509± 0.0937	0.0263± 0.0133
$\hat{c}_{ge}^2$	0.3281	0.1647	0.2839
($\hat{h}_{ml}^2$)	0.5909	0.8308	0.4644
($\hat{\sigma}_g^2$)	8141.9579	19858.7067	1688.7897
($\hat{\sigma}_c^2$)	16256.8161	9320.5998	18173.7592
($\hat{\sigma}_e^2$)	25152.9360	27399.7351	44158.01158
($\hat{\sigma}_f^2$)	49551.7101	56579.0418	64020.5605
$\hat{r}_{gloc}$	0.3337	0.6805	0.0850
$\hat{r}_{g\hat{g}}$	0.7687	0.9115	0.6815
Média geral	1414.0093	615.4906	1152.7441
(CV_g%)	6.3813	22.8957	3.5649
(CV_e%)	11.2161	26.8937	18.2293

Verifica-se, pela Tabela 72, a presença de variabilidade genética significativa entre as linhagens em avaliação, conforme demonstrado pelas estimativas de herdabilidade e seus desvios padrões. Os referidos desvios padrões foram de pequena magnitude em comparação com as magnitudes das herdabilidades, garantindo que as mesmas não atingiram o valor zero, via os limites inferiores dos intervalos de confiança (dados por aproximadamente –1,96 vezes os desvios padrões), fato que denotaria ausência de variabilidade genética.

Os coeficientes de variação genotípica apresentaram valores baixos, exceto para o local 2. Entretanto, as estimativas de herdabilidades em nível de médias de linhagens apresentaram valores moderados a altos (0,46 a 0,83), conduzindo a ótimas (0,68 a 0,91) acurácias na seleção de linhagens (Tabela 72). A correlação genotípica da performance das linhagens através dos 15 ambientes equivaleu a 0,085, revelando um alto nível de interação genótipos x ambientes do tipo complexa, mostrando que os melhores linhagens em um ambiente não necessariamente serão as melhores em outros. Isto justifica considerar na seleção das linhagens, suas estabilidades e adaptabilidades.

As estimativas de herdabilidades no sentido amplo, em nível individual, apresentaram valores variando de 0,03 a 0,35 (Tabela 72), apresentando o menor valor na análise conjunta dos 15 ambientes, devido à consideração dos efeitos da interação envolvendo simultaneamente genótipos, anos e épocas. Essa herdabilidade não está inflacionada pela interação genótipos x ambientes, ou seja, é livre de todas essas interações. Os coeficientes de determinação dos efeitos da interação genótipos x ambientes variaram de 16,5 % a 32,8 %. Esses valores referem-se à proporção da variabilidade fenotípica total explicada pela interação.

Os coeficientes de variação experimental apresentaram moderadas magnitudes, revelando boa precisão experimental. Em conjunto com o nível de variação genotípica e com o número de repetições empregado, foram suficientes para propiciar altas acurácias seletivas (68 % a 91 %). Maiores detalhes sobre acurácia e relações com o número de repetições e coeficientes de variação genotípica e experimental foram apresentados no Capítulo 3.

4.2 Seleção Genotípica por Local e para o Conjunto de Ambientes

Na Tabela 73 são apresentados o ordenamento e valores genotípicos preditos das 20 linhagens e ganhos genéticos estimados com a seleção das cinco melhores, para o caráter produção, em cada um dos dois locais e também para a análise conjunta dos 15 ambientes. Para um ambiente médio representado pelos 15 experimentos, as cinco melhores linhagens foram 1, 20, 4, 15 e 8. Ganhos genéticos da ordem de 4 % podem ser obtidos com a seleção das duas melhores linhagens.

No primeiro local, as cinco melhores linhagens foram 15, 4, 19, 16 e 8. No segundo local, as cinco melhores linhagens foram 1, 20, 14, 7 e 11. Dentre as cinco melhores, pelo menos duas linhagens em cada local são coincidentes com os melhores na média dos três locais. Esta seleção por local usa simultaneamente a informação de todas as épocas e anos, e não apenas a informação de cada local, conforme usado tradicionalmente.

Tabela 73. Valores genotípicos de 20 genótipos de feijão em estudo e ganhos genéticos preditos dos cinco melhores para o caráter produção, em dois locais (Capão Bonito e Masul) e também na análise conjunta dos 15 ambientes, no Estado de São Paulo.

Linhagens	Local 1 (Capão Bonito)		Linhagens	Local 2 (Masul)		Linhagens	Análise Conjunta (15 Ambientes)	
	Valores genotípicos	Ganho genético		Valores genotípicos	Ganho genético		Valores genéticos	Ganho genético
	u_1 + g+ ge	(%)		u_2 + g+ ge	(%)		(u + g)	(%)
15	1536.5880	8.67	1	936.1611	52.10	1	1206.3723	4.65
4	1508.7766	6.70	20	842.1174	36.82	20	1202.6719	4.33
19	1491.7870	5.50	14	820.9303	33.38	4	1190.5022	3.28
16	1477.7520	4.51	7	690.1726	12.13	15	1180.7170	2.43
8	1477.5673	4.49	11	667.0641	8.38	8	1174.0206	1.85
1	1475.7206		9	661.3000		10	1165.5746	
11	1449.4975		3	638.1395		14	1160.4032	
12	1444.1420		4	634.0371		3	1156.3776	
6	1424.0130		8	629.1039		11	1156.2228	
20	1421.6123		10	603.6066		6	1155.2938	
5	1419.5809		15	588.4433		19	1154.1094	
18	1406.0999		6	540.7723		7	1147.6994	
3	1397.0511		18	538.9548		12	1146.3601	
2	1383.3855		16	527.9977		2	1128.5005	
10	1357.7163		12	523.0125		5	1128.1753	
17	1329.4618		5	517.8196		16	1127.0451	
9	1322.9983		17	514.0288		9	1122.7872	
13	1321.8903		13	507.5376		13	1122.4156	
7	1318.1969		19	465.8385		18	1115.3089	
14	1316.3502		2	462.7747		17	1114.3258	

4.3 Seleção Conjunta para Produtividade, Estabilidade e Adaptabilidade

Para a análise de estabilidade e adaptabilidade, atualmente, procedimentos de interpretação mais simples têm tido apelo. Neste sentido, medidas que incorporam ambos (estabilidade e adaptabilidade, juntamente com a produtividade) em uma única estatística, tais quais os métodos de Annicchiarico (1992) e Lin e Binns (1988) e modificações, têm sido enfatizados (Cruz e Carneiro, 2003). No contexto dos modelos mistos, um método para ordenamento de genótipos simultaneamente por seus valores genéticos (produtividade) e estabilidade, refere-se ao procedimento BLUP sob médias harmônicas (Resende, 2002, p. 344). Quanto menor for o desvio padrão do comportamento genotípico através dos locais, maior será a média harmônica de seus valores genotípicos através dos locais. Assim, a seleção pelos maiores valores da média harmônica dos valores genotípicos (MHVG) implica simultaneamente seleção para produtividade e estabilidade.

Em termos de adaptabilidade, uma medida simples e eficiente no contexto dos modelos mistos refere-se à performance relativa dos valores genotípicos (PRVG) através dos ambientes. Neste caso, os valores genotípicos preditos (ou os dados originais) são expressos como proporção da média geral de cada local e, posteriormente, obtém-se o valor médio desta proporção através dos locais. Genericamente, a performance relativa tem sido usada há longo tempo (Wright et al. 1966) em termos de dados fenotípicos e constitui a base do método de Annicchiarico (1992).

A seleção simultaneamente por produtividade, estabilidade e adaptabilidade, no contexto dos modelos mistos, pode ser realizada pelo método da média harmônica da performance relativa dos valores genéticos (MHPRVG) preditos. Este método permite selecionar simultaneamente pelos três atributos mencionados e apresenta as seguintes vantagens: (i) considera os efeitos genotípicos como aleatórios e portanto fornece estabilidade e adaptabilidade genotípica e não fenotípica; (ii) permite lidar com desbalanceamento; (iii) permite lidar com delineamentos não ortogonais; (iv) permite lidar com heterogeneidade de variâncias; (v) permite considerar erros correlacionados dentro de locais; (vi) fornece valores genéticos já descontados (penalizados) da instabilidade; (vii) pode ser aplicado com qualquer número de ambientes; (viii) permite considerar a estabilidade e adaptabilidade na seleção de indivíduos dentro de progênie; (ix) não depende da estimação de

outros parâmetros tais quais coeficientes de regressão; (x) elimina os ruídos da interação genótipos x ambientes pois considera a herdabilidade desses efeitos; (xi) gera resultados na própria grandeza ou escala do caráter avaliado; (xii) permite computar o ganho genético com a seleção pelos três atributos simultaneamente. Estes últimos dois fatores são bastante importantes. Outros métodos como o de Lin e Binns fornecem resultados que não são interpretados diretamente como valores genéticos e, portanto, não permitem computar o ganho genético no caráter composto pela produtividade, estabilidade e adaptabilidade.

Na Tabela 74, estão apresentados os resultados sobre a estabilidade – Média Harmônica dos Valores Genotípicos através dos locais (MHVG) - adaptabilidade - Performance Relativa dos Valores Genotípicos em relação a média de cada local (PRVG) - e estabilidade e adaptabilidade simultaneamente – Média Harmônica da Performance Relativa dos Valores Genotípicos (MHPRVG) - para o caráter produção.

Tabela 74. Estabilidade de valores genotípicos (MHVG), adaptabilidade de valores genotípicos (PRVG), estabilidade e adaptabilidade de valores genotípicos (MHPRVG) para a produção de linhagens de feijoeiro.

Genótipo	MHVG	Genótipo	PRVG	PRVG*MG	Genótipo	MHPRVG	MHPRVG*MG
1	990.706	1	1.1653	1343.3279	20	1.1063	1275.3018
20	981.081	20	1.1472	1322.3826	1	1.0989	1266.6990
14	915.302	14	1.0756	1239.8577	4	1.0587	1220.3532
4	844.955	4	1.0614	1223.5172	14	1.0377	1196.2578
8	836.529	8	1.0395	1198.3090	8	1.0376	1196.0412
11	806.334	11	1.0236	1179.9209	11	1.0172	1172.5349
9	790.539	15	1.0197	1175.4173	15	1.0135	1168.3033
10	787.751	10	1.0149	1169.9164	10	1.0072	1161.0412
3	779.256	3	1.0017	1154.7055	3	0.9991	1151.7352
15	768.321	7	0.9947	1146.6883	12	0.9762	1125.3549
12	751.914	12	0.9835	1133.6852	7	0.9718	1120.2298
5	733.841	9	0.9718	1120.2215	9	0.9586	1105.0566
7	694.451	6	0.9706	1118.8854	6	0.9482	1093.0425
18	679.719	19	0.9530	1098.5454	5	0.9432	1087.2484
17	660.871	5	0.9515	1096.8694	18	0.9172	1057.2641
16	647.363	16	0.9332	1075.7608	16	0.9170	1057.0970
6	646.787	2	0.9328	1075.2870	19	0.9169	1056.9105
2	636.225	13	0.9242	1065.3159	2	0.9072	1045.7381
19	619.928	18	0.9241	1065.3074	17	0.8997	1037.1716
13	596.889	17	0.9117	1050.9627	13	0.8954	1032.1951

Verifica-se (Tabela 74) que as cinco melhores linhagens, com base nos critérios PRVG, MHVG e MHPRVG não são exatamente as cinco melhores pelo critério de produtividade média (Tabela 73). A coincidência foi de 80 % dentre as cinco melhores e houve inversão de ordem dentre os coincidentes. Isto mostra que a utilização desses novos atributos ou critérios de seleção podem propiciar um refinamento a mais na seleção. As duas melhores linhagens pelo critério MHPRVG apresentaram superioridade média de 10 % sobre a média geral dos 15 ambientes. Estes valores foram computados já penalizando as linhagens pela instabilidade através dos locais e ao mesmo

tempo capitalizando a capacidade de resposta (adaptabilidade) à melhoria do ambiente. Essas propriedades são intrínsecas ao método MHPRVG. Os valores de PRVG e MHPRVG, na Tabela 74, indicam exatamente a superioridade média do genótipo em relação à média do ambiente em que for cultivado. Assim, o genótipo 20 responde em média 1,1063 vezes a média do ambiente em que for plantado. O valor de MHPRVG*MG fornece o valor genotípico médio das linhagens nos locais avaliados, valor este já penalizado pela instabilidade e capitalizado pela adaptabilidade.

As três melhores linhagens a serem selecionados com base no método da MHPRVG são 20, 1 e 4. Tal seleção propicia um ganho de 8,7 % sobre a média geral dos 15 ambientes, considerando simultaneamente a produtividade, estabilidade e adaptabilidade através dos locais, épocas e anos.

Na Tabela 75 são apresentados os resultados referentes à seleção simultânea por produtividade, adaptabilidade e estabilidade por meio do emprego do método de Lin e Binns (1988) sobre os valores genotípicos preditos.

Tabela 75. Estabilidade e adaptabilidade de valores genotípicos para produção por meio do método (Pi) de Lin e Binns (1988), em que $Pi = \sum_j (VG_{ij} - M_j)^2/(2L)$, VG_{ij} é o valor genotípico do genótipo i no local j, M_j é o valor genotípico máximo no local j e L é o número de locais.

Genótipo	*Pi*
1	12058.32
20	15329.12
4	16914.68
8	21494.26
15	23792.42
10	28425.72
3	28913.15
14	28983.4
6	30563.59
11	30964.42
7	36578.01
12	37603.78
19	40882.28
9	41793.21
5	42302.30
16	44697.89
13	45364.89
2	47175.70
18	48533.96
17	49773.91

Verifica-se que dentre as dez melhores linhagens selecionadas pela MHPRVG, nove coincidem com as dez linhagens selecionadas pelo método de Lin e Binns (nesse método, os melhores materiais genéticos são aqueles com menores valores da estatística Pi). E tomando por base o método MHPRVG, o genótipo não coincidente foi o 12, o qual ficou em décimo lugar no ordenamento por esse método e em décimo segundo no método de Lin e Binns, portanto, em posições muito próximas. As três melhores linhagens foram as mesmas pelos dois métodos. A

correlação estimada entre os parâmetros dos dois métodos foi de alta magnitude (–0,9143). O método MHPRVG foi computado em referência à média do ambiente e não em relação ao melhor genótipo em cada ambiente como foi realizado pelo método de Linn e Binns. Poderia ter sido computado dessa forma e a concordância entre os dois métodos poderia ser ainda maior. O método de Annicchiarico foi também computado e apresentou correlação absoluta nessa mesma magnitude (porém positiva) com o método MHPRVG. Isso confirma que os três métodos utilizam basicamente os mesmos princípios e conceitos e praticamente selecionam os mesmos genótipos. O método MHPRVG apresenta a vantagem de fornecer resultados na própria escala de medição do caráter, os quais podem ser interpretados diretamente como valores genéticos para o caráter avaliado. Isto permite também calcular o ganho genético com a seleção simultânea para produtividade, adaptabilidade e estabilidade. Isto não é possível com o método de Lin e Bins. Assim, a estatística MHPRVG pode ser usada vantajosamente no contexto dos modelos mistos com efeitos genéticos aleatórios. A consideração dos efeitos genéticos e da interação g x e como aleatórios propicia vantagem também sobre o método AMMI (Gauch, 1988), o qual trata esses efeitos como fixos e portanto atua no nível fenotípico e não genotípico. É importante relatar que o BLUP dos efeitos da interação já elimina os ruídos de tais efeitos, à semelhança do método AMMI, conforme relatado por Resende (2004) e Resende e Thompson (2004). O método MHPRVG pode também ser aplicado separadamente para duas classes de ambientes: com índice ambiental negativo e positivo. Isto foi realizado e os resultados são apresentados na Tabela 76.

Tabela 76. Estabilidade de valores genotípicos (MHVG), adaptabilidade de valores genotípicos (PRVG), estabilidade e adaptabilidade de valores genotípicos (MHPRVG) para a produção de linhagens de feijoeiro, em ambientes desfavoráveis e favoráveis.

Desfavoráveis				Favoráveis			
Ordem	**Genótipo**	**MHPRVG**	**MHPRVG*MG**	**Ordem**	**Genótipo**	**MHPRVG**	**MHPRVG*MG**
1	20	1.1814	883.6236	1	15	1.0551	1856.9155
2	1	1.1561	864.7041	2	4	1.0480	1844.4361
3	14	1.1454	856.7195	3	10	1.0382	1827.2043
4	4	1.0583	791.5611	4	8	1.0381	1827.0434
5	11	1.0547	788.8949	5	1	1.0282	1809.6280
6	8	1.0290	769.6291	6	20	1.0214	1797.6532
7	7	1.0092	754.8332	7	19	1.0132	1783.0739
8	3	1.0004	748.2888	8	6	1.0093	1776.2871
9	9	0.9920	741.9386	9	5	1.0062	1770.9233
10	10	0.9700	725.5403	10	3	0.9972	1754.9216
11	12	0.9648	721.6188	11	12	0.9944	1750.0559
12	15	0.9631	720.3304	12	2	0.9925	1746.7914
13	18	0.9002	673.3250	13	11	0.9826	1729.3404
14	5	0.8927	667.6885	14	16	0.9823	1728.7816
15	6	0.8919	667.0914	15	13	0.9707	1708.4199
16	16	0.8651	647.0836	16	7	0.9558	1682.0996
17	17	0.8638	646.0955	17	17	0.9529	1677.0075
18	13	0.8479	634.2073	18	18	0.9502	1672.3484
19	2	0.8462	632.8997	19	14	0.9382	1651.2122
20	19	0.8253	617.3202	20	9	0.9313	1638.9397

Verifica-se que o ganho genético com a seleção pelos três atributos simultaneamente é maior nos ambientes desfavoráveis (2, 5, 6, 9, 10, 11,12, 13 e 14) do que nos favoráveis (1, 3, 4, 7, 8 e 15). O ganho com a seleção dos três melhores genótipos nos ambientes desfavoráveis foi acima de 15 % em relação à média dos nove ambientes desfavoráveis. Este valor é obtido fazendo-se as médias dos três primeiros valores na coluna MHPRVG e verificando o quanto esse resultado é superior a 1 (Tabela 76).

Nos ambientes favoráveis, esse ganho foi acima de 4 %. Dentre os seis melhores, quatro foram coincidentes nos ambientes favoráveis e desfavoráveis (linhagens 20, 1, 4 e 8). Dentre esses seis, duas inversões importantes foram observadas: as linhagens 15 e 10 foram a primeira e a terceira colocadas nos ambientes favoráveis e apenas a décima segunda e décima nos desfavoráveis, respectivamente (Tabela 76).

Em resumo, os métodos que mais penalizam os valores genotípicos preditos são, pela ordem: MHVG, MHPRVG e PRVG (Tabela 74). De maneira genérica, pode-se dizer que os métodos MHVG e MHPRVG são opções seguras, sendo o MHVG um pouco mais conservador. As linhagens 20 e 1 foram as primeiras colocadas em todos os critérios: produtividade, estabilidade e adaptabilidade e os três atributos simultaneamente em todos os ambientes. Nas demais posições houve uma certa alternância de genótipos de acordo com o critério (Tabela 74).

Também no contexto dos modelos mistos com efeitos aleatórios de genótipos, uma abordagem mais completa e detalhada da natureza da interação refere-se à técnica FAMM relatada por Smith et al. (2001) e Resende e Thompson (2003; 2004) e apresentada em detalhes no Capítulo 8.

CAPÍTULO 11

SELEÇÃO GENÔMICA AMPLA (GWS)

E MODELOS LINEARES MISTOS

1 FUNDAMENTOS DA *GENOME WIDE SELECTION* (GWS)

A seleção genética tem sido praticada pelo procedimento BLUP (em suas versões frequentista e bayesiana) usando dados fenotípicos avaliados a campo. Uma primeira proposição realizada para aumentar a eficiência desse procedimento baseado em dados fenotípicos foi descrita por Lande & Thompson (1990), por meio da seleção auxiliada por marcadores (MAS) moleculares. A MAS utiliza simultaneamente dados fenotípicos e dados de marcadores moleculares em ligação gênica próxima com alguns locos controladores de características quantitativas (QTL). Em geral, os dados de marcadores são utilizados como covariáveis na explicação dos valores fenotípicos dos indivíduos em avaliação ou como efeitos aleatórios incorporados no modelo para o fenótipo. Esses marcadores são eleitos ou não como determinantes dos efeitos de QTLs após modelagem

estatística associada a erros do tipo II (probabilidade de aceitar uma hipótese falsa, ou seja, tomar como verdadeira uma hipótese falsa de ausência de efeitos).

A seleção baseada na MAS apresenta as seguintes características:

- requer o estabelecimento (análise de ligação) de associações marcadores-QTLs para cada família em avaliação, ou seja, essas associações apresentam utilidade para seleção apenas dentro de cada família mapeada em espécies alógamas.

- para ser útil, precisa explicar grande parte da variação genética de uma característica quantitativa, que é governada por muitos locos de pequenos efeitos. Isto não tem sido observado na prática, exatamente em função da natureza poligênica e alta influência ambiental nos caracteres quantitativos, fato que conduz à detecção apenas de um pequeno número de QTLs de grandes efeitos, os quais não explicam suficientemente toda a variação genética.

- só apresenta superioridade considerável em relação à seleção baseada em dados fenotípicos, quando o tamanho de família avaliado e genotipado é muito grande (da ordem de 500 ou mais).

Em função desses aspectos, a implementação da MAS tem sido limitada e os ganhos em eficiência muito reduzidos (Dekkers, 2004).

O grande atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações de DNA na seleção, de forma a permitir alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos. Visando a esses objetivos, Meuwissen et al. (2001) propuseram um novo método de seleção denominado seleção genômica (GS) ou seleção genômica ampla (*genome wide selection* – GWS), a qual pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético de espécies alógamas, apresenta alta acurácia seletiva para a seleção baseada exclusivamente em marcadores e não exige prévio conhecimento das posições (mapa) dos QTLs, não estando sujeita aos erros tipo II associados à seleção de marcadores ligados a QTLs.

Esse método permaneceu discreto por cerca de cinco anos, devido ao fato dos marcadores moleculares disponíveis à época serem caros e restritos. Recentemente, com o desenvolvimento e baixo custo dos marcadores tipo SNP (*single nucleotide polymorphism*), o método tornou-se muito

atrativo e geneticistas e melhoristas renomados e adeptos de métodos tradicionais têm demonstrado e confirmado a superioridade e exeqüibilidade prática do método em benefício do melhoramento animal (Schaeffer, 2006; Kolbehdari et al. 2007; Meuwissen, 2007; Goddard & Hayes, 2007; Long et al. 2007; Legarra & Misztal, 2008) e vegetal (Bernardo, 2007). Esses trabalhos mostraram, definitivamente, que a seleção genômica terá grande utilidade no melhoramento genético, via métodos do tipo BLUP/GWS, que equivalem ao procedimento BLUP aplicado sobre dados moleculares e permitem a predição de valores genéticos genômicos. A GWS é excelente para caracteres de baixa herdabilidade, ao contrário da MAS, que não é útil para caracteres de baixa herdabilidade (Muir, 2007).

A MAS baseia-se na detecção, mapeamento e uso de QTLs na seleção. Ou seja, enfatiza a determinação do número, posição e efeitos dos QTLs marcados. A GWS é definida como a seleção simultânea para centenas ou milhares de marcadores, os quais cobrem o genoma de uma maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Esses marcadores em desequilíbrio de ligação com os QTL´s, tanto de grandes quanto de pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo. Por probabilidade, cada QTL estará em desequilíbrio de ligação com pelo menos um marcador. Somente os marcadores em desequilíbrio de ligação com os QTL´s é que explicarão os fenótipos e a variação genética. Os efeitos dos marcadores são estimados em uma amostra de indivíduos pertencentes a várias famílias. Assim, o impacto de determinadas famílias específicas (com específicos padrões de desequilíbrio de ligação) nas estimativas dos efeitos dos marcadores será minimizado.

A GWS é ampla porque atua em todo o genoma sem a necessidade prévia de identificar os marcadores com efeitos significativos e de mapear QTLs. Valores genéticos genômicos associados a cada marcador ou alelo são usados para fornecer o valor genético genômico global de cada indivíduo. Há uma diferença básica na predição de valores genéticos tradicionais e na predição de valores genéticos genômicos. Nos primeiros, informações fenotípicas são utilizadas visando inferências sobre os efeitos dos genótipos dos indivíduos e, nos últimos, informações genotípicas (genótipos para os alelos marcadores) são usadas visando inferências sobre os valores fenotípicos futuros (ou valores genéticos genômicos preditos) dos indivíduos. Em outras palavras, os métodos

tradicionais usam o fenótipo para inferir sobre o efeito do genótipo e a GWS usa o genótipo para inferir sobre o fenótipo a ser expresso.

A GWS pode basear-se no uso de: (i) apenas dos marcadores; (ii) de haplótipos ou intervalos definidos por 2 marcadores; (iii) haplótipos definidos por mais de 2 marcadores, incluindo a covariância entre haplótipos devida à ligação. Segundo Callus et al. (2008), para caracteres de baixa herdabilidade (10%) não existem diferenças significativas entre essas 3 abordagens. Solberg et al. (2006) mostraram que é possível praticar a GWS eficientemente com o uso apenas dos marcadores, ou seja, com a predição direta dos efeitos dos marcadores. Relatam também que isso é vantajoso porque não há necessidade de estimar as fases de ligação entre os marcadores, as quais são estimadas com algum erro. Não apenas marcadores SNPs podem ser usados na GWS. Marcadores microssatélites também se prestam a esse fim. Solberg et al. (2006) relatam que o uso de SNPs requer 4 a 5 vezes maior densidade de marcadores do que o uso de microssatélites. Isto se deve à natureza bi-alélica dos SNPs e multi-alélica dos microssatélites.

A GWS fundamenta-se nos marcadores genéticos moleculares do tipo SNP (polimorfismo de um único nucleotídeo), os quais baseiam-se na detecção de polimorfismos resultantes da alteração de uma única base no genoma. E para que uma variação seja considerada SNP, essa deve ocorrer em pelo menos 1 % da população. Os SNPs são a forma mais abundante de variação do DNA em genomas e são preferidos em relação a outros marcadores genéticos devido à sua baixa taxa de mutação e facilidade de genotipagem. Milhares de SNPs podem ser usados para cobrir o genoma de um organismo com marcadores que não estão a mais de 1 cM um do outro no genoma inteiro.

A GWS atua mais proximamente aos QTNs (nucleotídeos de características quantitativas) ou sobre marcadores fortemente ligados a esses. Os QTNs são polimorfismos funcionais, causadores diretos da variação quantitativa observada. A análise de SNP´s permite a detecção de polimorfismos funcionais ou polimorfismos em forte desequilíbrio de ligação com os QTNs. Tecnologias para genotipagem de milhares de SNPs em microarranjos estão disponíveis atualmente. Microarranjos são sistemas de arranjos de DNA que utilizam lâminas de vidro e sondas fluorescentes e permitem depositar milhares de seqüências de DNA. Nessa técnica são utilizados nucleotídeos marcados capazes de emitir fluorescência ao invés de radioatividade.

O desenvolvimento teórico da GWS coincide com a tecnologia SNP, a qual é acurada e relativamente barata. A GWS usa associações de um grande número de marcadores SNPs em todo o genoma com os fenótipos, capitalizando no desequilíbrio de ligação entre os marcadores e QTLs proximamente ligados, sem uma prévia escolha de marcadores com base nas significâncias de suas associações com o fenótipo. Predições são então obtidas para os efeitos dos haplótipos marcadores ou dos alelos em cada marcador. Essas predições derivadas de dados fenotípicos e de genótipos SNPs em alta densidade em uma geração são então usadas para obtenção dos valores genéticos genômicos (VGG) dos indivíduos de qualquer geração subseqüente, tendo por base os seus próprios genótipos marcadores. Os hapótipos são definidos como intervalos resultantes de combinações de dois alelos marcadores vizinhos. A seleção genômica baseada simultaneamente em um grande número de marcadores contrasta com a MAS, que é baseada em um número limitado de marcadores ou genes. Os marcadores moleculares do tipo microssatélites podem também ser usados na GWS. Tais marcadores são eficientes por serem co-dominantes, multi-alélicos, abundantes e apresentarem alta transferibilidade entre indivíduos e espécies.

Se os marcadores estão ligados aos QTLs e não aos nucleotídeos causadores da variação quantitativa (QTNs), variantes das fases de ligação fazem com que os marcadores sejam incorretos em algumas famílias ou populações. Com o uso dos SNPs, existe a vantagem de que os mesmos tendem a estar intimamente ligados aos próprios QTNs.

Quando o desequilíbrio de ligação entre marcadores não é completo, as freqüências gênicas podem mudar substancialmente através das gerações. Também, se houver dominância e epistasia em magnitudes consideráveis, os efeitos dos marcadores necessitarão ser re-estimados para manter a acurácia da GWS em várias gerações (Dekkers, 2007). O desequilíbrio de ligação ou desequilíbrio de fase gamética é uma medida da dependência ou não entre alelos de dois ou mais locos. Em um grupo de indivíduos, se dois alelos são encontrados juntos com freqüência maior do que aquela esperada com base no produto de suas freqüências, infere-se que tais alelos estão em desequilíbrio de ligação. Valores de desequilíbrio de ligação próximos de zero indicam equilíbrio ou independência entre os alelos de diferentes genes e valores próximos de um indicam desequilíbrio ou ligação entre alelos de diferentes genes.

Com desequilíbrio de ligação completo e ausência de dominância e epistasia, os VGG são caracteres genéticos, com herdabilidade 1 e cujos efeitos permanecem constantes através das gerações. Embora sejam estimativas, esses caracteres genéticos podem ser vistos como herdados de maneira poligênica, porém sem efeitos ambientais. Os VGG usam valores genéticos de cada alelo e de cada gene (loco) do caráter quantitativo.

A acurácia da seleção GWS depende da proporção (m^2) da variação genética explicada pelos marcadores e da acurácia da predição dos efeitos dos haplótipos marcadores que estão em desequilíbrio de ligação com os QTL´s ($r_{qq'}$). O parâmetro m^2 depende da densidade de marcadores e da extensão e padrão do desequilíbrio de ligação que existe na população. Por sua vez, o parâmetro $r_{qq'}$ depende da quantidade e precisão dos dados disponíveis para estimar os efeitos dos marcadores e da eficiência da estratégia e dos métodos estatísticos usados na predição.

2 PROCEDIMENTO REML/BLUP/GWS

A estimação dos VGGs usa um conjunto de dados de referência que inclui indivíduos com ambos conhecidos, os genótipos (marcadores) e os fenótipos. Os valores genéticos estimados dos haplótipos em um grande número de supostos caracteres quantitativos são usados para a predição dos valores genéticos genômicos de indivíduos jovens canditatos à seleção e que foram genotipados para os marcadores mas não possuem informação fenotípica. Se toda variação genética puder ser explicada pelos haplótipos, não há necessidade de inclusão no modelo de predição, do efeito poligênico para levar em consideração a variação genética não explicada pelos haplótipos (variação genética residual). Na prática, se não há uma cobertura completa (mapa denso de marcadores) do genoma com SNPs, a inclusão do efeito poligênico pode tornar-se necessária.

A estimação dos valores genéticos genômicos para haplótipos marcadores individuais ou alelos individuais do QTL baseia-se em um relativamente grande número de haplótipos e um relativamente pequeno número de indivíduos. Marcadores SNPs circundando cada região genômica de 1 cM são combinados em um haplótipo marcador. Com um mapa de marcadores denso, alguns marcadores estarão muito próximos dos QTLs e provavelmente em desequilíbrio de ligação com

eles. Assim, alguns alelos marcadores estarão correlacionados com efeitos positivos no caráter quantitativo através de todas as famílias e poderão ser usados na seleção sem a necessidade de estabelecer a fase de ligação em cada família. Segmentos cromossômicos que contém os mesmos haplótipos marcadores raros apresentam alta probabilidade de identidade por descendência e então carregam os mesmos alelos do QTL. A precisão do mapeamento de QTL pelos métodos tradicionais de análise de ligação é pouco melhorada pelo uso de mapas de marcadores densos. Mas, pela abordagem da GWS, os efeitos nos QTLs, de pequenos segmentos de cromossomo definidos pelos haplótipos dos alelos marcadores que eles carregam são estimados com alta precisão e os referidos mapas densos são muito úteis (Meuwissen et al., 2001).

Para um genoma de 3.000 cM apenas 3.001 marcadores a intervalos de 1 cM seriam necessários. No entanto, tais marcadores necessitam ser informativos e um painel com 10.000 marcadores aumentaria as chances de sucesso. Cada par contíguo de marcadores define um haplótipo ou intervalo. Existem apenas dois alelos para cada marcador, pois os SNPs têm diferenças em um único par de bases. Dessa forma, para cada par de marcadores existem quatro haplótipos possíveis. A freqüência de cada haplótipo depende da freqüência dos alelos em cada marcador e a distância entre marcadores depende dos eventos de recombinação. Assim, um número suficiente de indivíduos devem ser genotipados de forma que todos os haplótipos estejam representados nos indivíduos com avaliações fenotípicas (Schaeffer, 2006).

A estimação dos efeitos do elevado número de haplótipos a partir de um número limitado de dados conduz ao problema da estimação por quadrados mínimos com insuficiente número de graus de liberdade para ajustar todos esses efeitos simultaneamente. O método BLUP por outro lado, permite ajustar todos os efeitos alélicos simultaneamente mesmo quando existem mais efeitos a serem preditos do que o número total de observações fenotípicas.

Os efeitos estimados dos haplótipos são assumidos como estimativas válidas para toda a população e não para apenas um grupo de indivíduos. Dessa forma, VGG podem ser estimados para quaisquer indivíduos da população desde que os mesmos sejam genotipados e os haplótipos marcadores sejam determinados. Assim, cada indivíduo pode ter uma estimativa de VGG desde o momento em que é gerado.

Os efeitos de cada intervalo em um caráter em um dado ambiente podem ser estimados para todos os intervalos simultaneamente em um modelo linear misto em que os efeitos de intervalo são tratados como aleatórios. Os genótipos marcadores dos indivíduos podem ser usados para predição de qualquer caráter, mas as estimativas dos efeitos dos intervalos ou haplótipos serão diferentes para cada caráter. Os intervalos com maiores efeitos em cada caráter conterão um ou mais QTLs. A maioria dos intervalos, no entanto, apresentará efeitos relativamente menores, refletindo o que acontece no modelo infinitesimal (muitos genes de pequenos efeitos associados ao caráter quantitativo).

Para uso dos SNPs na GWS, inicialmente devem ser identificados aqueles informativos e, posteriormente, um *software* deve ser usado para construir haplótipos a partir dos genótipos SNP. De posse dos haplótipos, as predições de seus efeitos podem ser feitas por meio de *softwares* específicos de genética quantitativa e estatística.

Modelo Linear para os Haplótipos

O seguinte modelo linear misto geral é usado para estimar os efeitos de haplótipos:

$y = Xb + Zh + e,$

em que y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, h é o vetor dos efeitos aleatórios de haplótipos (intervalos) e e refere-se ao vetor de resíduos aleatórios. X e Z são as matrizes de incidência para b e h.

A dimensão de h é igual ao número de intervalos multiplicado por 4 (número de haplótipos possíveis para cada intervalo). A matriz de incidência Z contém os valores 0, 1 e 2 para o número de alelos (do suposto QTL) ou haplótipos do tipo h_i no indivíduo diplóide j.

A estrutura de médias e variâncias é definida como:

$$h \sim N(0,\ G) \qquad E(y) = Xb$$

$$e \sim N(0,\ R = I\sigma_e^2) \qquad Var(y) = V = ZGZ' + R$$

$$G = \sum_i^n I_h \sigma_i^2$$

em que I_h é de ordem 4 e σ_i^2 é a variância dos efeitos dos haplótipos no i-ésimo intervalo e n é o número total de intervalos.

As equações de modelo misto para a predição de h via o método BLUP/GWS equivalem a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$ em que σ_g^2 refere-se à variância genética total do caráter e σ_e^2 é a variância residual. O valor genético genômico global do indivíduo j é dado por $VGG = \hat{y}_j = \sum_i Z_i \hat{h}_i$,

em que Z_i equivale a 0, 1 ou 2.

As equações de predição apresentadas acima assumem *a priori* que todos os locos explicam iguais quantidades da variação genética. Assim, a variação genética explicada por cada loco é dada por σ_g^2 / n, em que σ_g^2 é a variação genética total e n é o número de intervalos ou haplótipos. Essa estratégia foi adotada por Meuwissen et al. (2001), Muir (2007), Bernardo (2007) e Kolbehdari et al. (2007). Bernardo (2007) relata que essa suposição de iguais variâncias por loco não conduz a perdas significativas na acurácia da GWS. A variação genotípica σ_g^2 pode ser estimada por REML sobre os dados fenotípicos da maneira tradicional ou pela própria variação entre os haplótipos ou variância dos segmentos cromossômicos de QTL.

Os efeitos do vetor h são ajustados como covariáveis aleatórias associadas às observações fenotípicas. Muir (2007) denomina o método BLUP/GWS como regressão de cumeeira ou "*ridge regression*" (RR). No caso, o parâmetro de regressão é função de $\sigma_e^2 / (\sigma_g^2 / n)$.

O procedimento BLUP/GWS é similar ao BLUP tradicional. Porém, na predição dos referidos efeitos aleatórios não há necessidade de uso da matriz de parentesco (Schaeffer, 2006). A matriz de parentesco baseada em pedigree usada no BLUP tradicional é substituída por uma matriz de parentesco estimada pelos marcadores. Essa matriz de parentesco é a própria matriz Z'Z presente nas equações de modelo misto, em que Z é a matriz de incidência para os efeitos de marcadores. Esse procedimento é superior ao uso do pedigree pois efetivamente captura a matriz de parentesco realizada e não uma matriz de parentesco médio associada ao pedigree.

Na situação em que os marcadores não explicam toda a variação genética, o modelo pode ser estendido para englobar o efeito poligênico residual (variação genética não explicada pelos marcadores). Esse modelo estendido é da forma

$y = Xb + Zh + Wg^* + e,$

em que g* é o vetor dos efeitos poligênicos residuais (aleatórios) e W é a matriz de incidência para g*.

Com o uso de mapa denso de marcadores a inclusão dos efeitos poligênicos g* não aumenta a acurácia da GWS (Calus & Veerkamp, 2007). No entanto, para capitalizar o ganho genético no longo prazo, a inclusão desses efeitos é recomendada (Muir, 2007). No longo prazo, o BLUP tradicional obtém informação no genoma inteiro em cada geração. A GWS sem o efeito poligênico seleciona de forma muito acurada para a mesma parte do genoma em cada geração. Uma forma de aliviar esse problema é por meio da re-estimação dos efeitos de marcadores freqüentemente, visando à exploração de novas associações marcadores-QTL.

Os modelos apresentados devem ser estendidos para a incorporação de outros fatores de efeitos aleatórios visando contemplar ajustes para efeitos ambientais tais quais efeitos de blocos incompletos e, também, para a incorporação de covariáveis ambientais de efeitos fixos. Podem ser estendidos também para incorporar efeitos de dominância, tais quais efeitos de capacidade específica de combinação. Os modelos apresentados assumem genes de efeitos aditivos e portanto estimam os efeitos médios dos genes.

Intervalos que contém QTL podem ser localizados por meio da soma dos efeitos absolutos dos haplótipos dentro de cada intervalo. Se não existe um QTL em um intervalo, todas as estimativas dos efeitos dos haplótipos dentro dele serão de pequena magnitude em módulo. Intervalos com as maiores soma desses efeitos absolutos provavelmente contém um QTL ou estarão adjacentes a um intervalo contendo um QTL. Dessa forma, a posição do QTL pode ser encontrada e a descoberta de QTL com grande efeito é facilitada. Adicionalmente, a MAS não será de fato necessária pois todos os QTLs poderão ser selecionados simultaneamente usando GWS. A alta acurácia da GWS conduzirá a grandes alterações nas estratégias de melhoramento em várias espécies.

3 PROCEDIMENTOS BAYESIANOS PARA A ESTIMAÇÃO DOS EFEITOS DE HAPLÓTIPOS

Vários métodos de predição de valores genéticos genômicos foram propostos: quadrados mínimos (LS), BLUP/GWS, BayesA e BayesB (Meuwissen et al., 2001) e aprendizado de máquina (AM de Long et al., 2007). Essas abordagens diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos; o AM assume que existe um número limitado de genes e de SNPs a serem ajustados; o método BayesB é intermediário entre esses dois, assumindo poucos genes de grandes efeitos e muitos genes com pequenos efeitos. No método Bayes B muitos efeitos de marcadores são assumidos como zero, *a priori*. Isso reduz o tamanho do genoma por meio da concentração nas partes do mesmo onde existem QTLs. O melhor método é aquele que reflete melhor a natureza biológica do caráter poligênico em questão, em termos de efeitos gênicos.

O método LS é ineficiente devido a: impossibilidade de estimar todos os efeitos simultaneamente, pois o número de efeitos a estimar é maior do que o número de dados; estimando um efeito de cada vez e verificando a sua significância, conduz a superestimativas dos efeitos significativos; a acurácia do método é baixa; somente QTLs de grande efeito serão detectados e usados e, conseqüentemente, nem toda a variação genética será capturada pelos marcadores.

O método LS assume distribuição *a priori* para os QTLs, com variância infinitamente grande, fato que é incompatível com a conhecida variância genética total. O BLUP/GWS assume os efeitos de QTL com distribuição normal com variância constante através dos segmentos cromossômicos. A distribuição dos efeitos de QTL é conhecida em poucos caracteres e espécies. Em gado bovino leiteiro, Goddard & Hayes (2007) relatam a presença de 150 QTLs para o caráter produção de leite e estimaram a distribuição de seus efeitos como aproximadamente exponencial. Com distribuição exponencial e não muitos efeitos com valor zero, o melhor estimador dos efeitos alélicos é denominado LASSO (Tibshirani, 1996). Entretanto, com muitos efeitos com valor zero, o LASSO não é adequado.

O método ideal de predição de valores genéticos genômicos equivale ao cálculo da média condicional do valor genético dado o genótipo do indivíduo em cada QTL. Essa média somente pode

ser calculada usando uma distribuição a priori dos efeitos dos QTLs. Considerando cada QTL em separado essa esperança condicional é dada por $\hat{h} = E(h|dados)$. O estimador apropriado segue o teorema de Bayes e é dado por $\hat{h} = \frac{\int h * f(m|h)\ f(h)\ d\ h}{\int f(m|h)\ f(h)\ d\ h}$, em que $f(m|h)$ é a verossimilhança dos dados (m) e $f(h)$ é a distribuição *a priori* dos efeitos dos QTLs. Esse estimador mostra que o método ideal depende da distribuição *a priori* dos efeitos de QTL. A presença de QTLs é testada em muitas posições (10.000 SNPs) e, portanto, não existe QTLs em muitas posições. Dessa forma, a distribuição *a priori* $f(h)$ deve ter uma alta probabilidade para $f(0)$. Para especificar essa alta probabilidade, deve-se ter uma noção de quantos QTLs controlam o caráter (Goddard & Hayes, 2007).

Nessa situação com muitos efeitos h iguais a zero, o método BLUP/GWS resulta em muitas estimativas de h próximas de zero, porém não iguais a zero. Na soma dessas estimativas, esse efeito acumulado introduz algum erro na predição. Os métodos bayesianos BayesA e BayesB relatados por Meuwissen et al. (2001) consideram mais adequadamente a distribuição *a priori* dos efeitos dos QTLs.

O método BayesA equivale ao método BLUP, porém as variâncias dos segmentos cromossômicos diferem para cada segmento e são estimadas sob esse modelo, considerando a informação combinada dos dados e da distribuição *a priori* para essas variâncias. Essa distribuição é tomada como uma qui-quadrado invertida e escalada. Para obtenção dessa informação combinada ou da distribuição *a posteriori* das variâncias, adota-se o procedimento da amostragem de Gibbs. Detalhes da estimação bayesiana são apresentados por Resende (2002) e Sorensen & Gianola (2002).

O método BayesB usa uma distribuição *a priori* dos efeitos dos QTLs com alta densidade em $\sigma_g^2 = 0$ e distribuição qui-quadrado invertida para $\sigma_g^2 > 0$. Assim, considera que em muitos locos não existe variação genética, ou seja, não estão segregando. A distribuição a priori do método BayesA não tem um pico de densidade em $\sigma_g^2 = 0$. Uma vez que não é possível uma amostragem de $\sigma_g^2 = 0$, o método da amostragem de Gibbs não pode ser usada no método Bayes B. Assim, o algoritmo de Metropolis-Hastings deve ser usado.

Para se ter uma noção do número de QTLs afetando um caráter quantitativo, as técnicas de mapeamento são importantes. Genericamente, essas técnicas podem envolver: análise de marcas simples; mapeamento por intervalo entre dois marcadores; mapeamento por intervalo composto (soma dos efeitos de intervalo); mapeamento por intervalo múltiplo (análise simultânea dos intervalos). Essa última tende a ser mais eficiente.

Quanto aos métodos, eles foram desenvolvidos principalmente para populações com dois alelos segregantes, tais quais populações de retrocruzamento, F_2 e linhagens homozigóticas recombinantes. Para progênies F_1 ou progênies de irmãos germanos de plantas alógamas, um único loco pode estar segregando para até quatro alelos, dois diferentes em cada genitor. Nessa situação, os métodos tradicionais não são eficientes. Um método muito eficiente nesse caso foi apresentado por Wu et al. (2002). Esse é um método de máxima verossimilhança que permite estimar as fases de ligação entre diferentes tipos de marcadores e a própria ligação, simultaneamente. Esse método, aplicado simultaneamente em um modelo para predição de valores genéticos por BLUP é recomendado para o mapeamento em famílias de irmãos germanos.

4 RELAÇÃO ENTRE BLUP TRADICIONAL E BLUP GENÔMICO

O efeito genético aditivo (a) de um indivíduo i, predito pelo BLUP tradicional, é dado por

$$\hat{a}_i = 0{,}5\ (\hat{a}_p + \hat{a}_m) + \hat{a}_d$$
$$= 0{,}5\ (\hat{a}_p + \hat{a}_m) + h_d^2\ (y - X\hat{b} - 0{,}5\ \hat{a}_p - 0{,}5\ \hat{a}_m)$$

em que $h_d^2 = (1/2\ \sigma_a^2)/(1/2\ \sigma_a^2 + \sigma_e^2)$ é a herdabilidade dentro de famílias de irmãos germanos, σ_a^2 é a variância genética aditiva e σ_e^2 é a variância residual. As demais quantidades são:

$\hat{a}_p$: efeito genético aditivo predito do genitor paterno.

$\hat{a}_m$: efeito genético aditivo predito do genitor materno.

$\hat{a}_d$: efeito genético aditivo predito do indivíduo dentro de família, ou seja, desvio em relação à média dos efeitos aditivos paterno e materno, explicado pela segregação de amostragem mendeliana que ocorre durante a formação de gametas.

Por esse procedimento, a fração $0{,}5\ (\hat{a}_p + \hat{a}_m)$ é predita com alta acurácia, pois, os efeitos aditivos dos genitores são preditos com base em várias repetições experimentais. Por outro lado, a quantidade $\hat{a}_d$ é predita com baixa acurácia devido ao fato de cada indivíduo ser único e, portanto, não propiciar repetições experimentais, exceto se for clonado. Adicionalmente, a variação dentro de famílias $(1/2\ \sigma_a^2)$ é considerada comum a todas as famílias, o que não é verdadeiro, exatamente devido às diferentes segregações associadas a cada família.

Para contornar esse problema, as seguintes medidas podem ser adotadas: (i) realizar teste clonal dos indivíduos e estimar diretamente o valor genético total do indivíduo; (ii) adotar o procedimento BLUP-VEG, conforme descrito no Capítulo 3, o qual considera variação dentro de família específica para cada família; (iii) adotar a MAS; (iv) adotar a GWS. Dentre essas alternativas, as melhores são o teste clonal e a GWS. O BLUP-VEG não permite repetições experimentais e não avalia diretamente (via DNA) as segregações realizadas. A MAS não abrange adequadamente todo o caráter poligênico.

Em resumo, todas essas técnicas visam melhorar a acurácia da estimativa $\hat{a}_d$, referente aos efeitos da segregação mendeliana. A GWS é o método que explora adequadamente a segregação de amostragem mendeliana que ocorre por ocasião da formação de gametas, pois captura a matriz de parentesco realizada e não uma matriz de parentesco médio associada ao pedigree. Uma vez que a GWS avalia diretamente o DNA associado (via marcadores) a cada loco de todo o caráter poligênico, avalia diretamente cada segregação em nível individual e não em nível médio. Assim, a GWS avalia diretamente o genótipo dos filhos, permite conhecer cada segregação e produz estimativas mais acuradas de valores genéticos por meio da melhor predição do termo referente à segregação mendeliana $\hat{a}_d$. Conforme Goddard & Hayes (2007), sob o modelo infinitesimal com grande número de locos de pequeno efeito, a GWS prediz os valores genéticos de maneira mais acurada do que o BLUP tradicional baseado em pedigree e dados fenotípicos. Bernardo (2007) relata a superioridade da GWS sobre a MAS.

A GWS enfatiza mais o termo referente à segregação mendeliana $\hat{a}_d$, dando mais peso a esse componente do que o faz o BLUP tradicional. Isso leva à seleção de menos indivíduos aparentados do que o faz o BLUP, reduzindo assim o incremento da endogamia na população. Daetwyler et al. (2007) relatam que a GWS aumenta em torno de 67 % a acurácia da predição de $\hat{a}_d$ em comparação com o BLUP tradicional e, conseqüentemente, elevou a acurácia da seleção individual de 71 % para 85 %.

No contexto da GWS, o modelo para o valor fenotípico y de um indivíduo é dado por $y = g + e = h + g^* + e$, em que g representa os efeitos genéticos, e refere-se aos efeitos ambientais, h refere-se aos efeitos genéticos explicados pelos marcadores e g* representa os efeitos genéticos residuais não explicados pelos marcadores. O efeito genético h de um indivíduo é estimado por meio de $\hat{h} = \sum_j (\hat{h}_j^p + \hat{h}_j^m)$, em que j refere-se a uma região genômica e $\hat{h}_j^p$ e $\hat{h}_j^m$ são as estimativas BLUP dos haplótipos marcadores paternais e maternais na região genômica j. O estimador $\hat{h}$ pode ser expresso alternativamente por $\hat{h} = 0{,}5\ (\hat{h}_p + \hat{h}_m) + [(\sum_j \hat{h}_j^p - 0{,}5\ \hat{h}_p) + (\sum_j \hat{h}_j^m - 0{,}5\ \hat{h}_m)] = \sum_j (\hat{h}_j^p + \hat{h}_j^m)$. Essa expressão é similar à expressão apresentada anteriormente para $\hat{a}_i$, sendo que o termo entre colchetes representa os efeitos da segregação mendeliana. No entanto, tal termo tem herdabilidade igual a 1 e não igual a h_d^2, pois o efeito genético genômico é condicional aos genótipos marcadores e previamente derivado de estimativas dos efeitos dos marcadores e, portanto, não contém efeito residual. Logicamente, isso assume que os haplótipos podem ser determinados sem erro (Dekkers, 2007). É importante relatar que a acurácia global da GWS não é 1, pois g é dado por h + g*. Ou seja, existe uma parte (g*) de g que não é explicada pelos marcadores e não é contemplada pela GWS.

5 IMPLEMENTAÇÃO DA SELEÇÃO GENÔMICA AMPLA

Na prática da seleção genômica ampla, três populações ou conjuntos de dados são necessários, conforme descrito na seqüência, com base em Goddard & Hayes (2007).

População de Descoberta. Esse conjunto de dados contempla um grande número de marcadores SNPs avaliados em um número moderado de indivíduos, os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Equações de predição de valores genéticos genômicos são obtidas para cada caráter de interesse. Essas equações associam a cada intervalo marcador o seu efeito (predito por BLUP) no caráter de interesse.

População de Validação. Esse conjunto de dados é menor do que aquele da população de descoberta e contempla indivíduos avaliados para os marcadores SNPs e para os vários caracteres de interesse. As equações de predição de valores genéticos genômicos são testadas para verificar suas acurácias nessa amostra independente. Para computar essa acurácia, os valores genéticos genômicos são preditos (usando os efeitos estimados na população de descoberta) e submetidos a análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos haplótipos marcadores, os erros dos valores genéticos genômicos e dos valores fenotípicos são independentes e toda correlação entre esses valores é de natureza genética e equivale à própria acurácia para a seleção individual.

População de Seleção. Esse conjunto de dados contempla apenas os marcadores SNPs avaliados nos candidatos à seleção. Essa população não necessita ter os seus fenótipos avaliados. As equações de predição derivadas na população de descoberta são então usadas na predição dos VGG ou fenótipos futuros dos candidatos à seleção. Mas a acurácia seletiva associada refere-se àquela calculada na população de validação.

Segundo Meuwissen (2007), quando dezenas a centenas de milhares de haplótipos são estimados, existe o risco de superparametrização, ou seja, erros nos dados serem explicados pelos efeitos de marcadores. A validação cruzada é então de grande importância para contornar esse problema.

Sob seleção genômica, todos os candidatos à seleção (indivíduos sem observação fenotípica) poderão ser avaliados para quaisquer ambientes, desde que tais ambientes possuam equações de predição derivadas para os próprios e com alta acurácia. Acurácia da ordem de 85 % para a GWS foi relatada por Meuwissen et al. (2001) para uma população com 2200 indivíduos com avaliações fenotípicas. Tais autores relataram também que equações de predição acuradas (71 %) foram obtidas mesmo para populações de descoberta de tamanho modesto, tal qual com 500 indivíduos

com avaliações fenotípicas. Assim, indivíduos poderão ser comercializados com base em seus valores fenotípicos preditos (valores de cultivo e uso – VCU), derivados de um catálogo de marcadores associados aos candidatos à seleção. Também os produtos animais e vegetais (leite, carne, alimentos, fibras) poderão ser remunerados com base em seus marcadores genéticos (Goddard & Hayes, 2007).

6 ASPECTOS COMPUTACIONAIS DA SELEÇÃO GENÔMICA AMPLA

O principal problema computacional da GWS é que a matriz de informação Z'Z muitas vezes não pode ser armazenada na memória RAM do computador, devido à sua grande dimensão. Assim, métodos iterativos de resolução das equações de modelo misto são necessários tal qual o método da iteração nos dados. O ajuste de milhares de covariáveis aleatórias implica densa matriz dos coeficientes das equações de modelo misto e grandes esforços e recursos computacionais. Legarra & Misztal (2007) indicam o método de Gauss-Seidel com atualização de resíduos como o mais apropriado para a predição BLUP e estimação de componentes de variância nesse contexto. O *software* Selegen-Genômica-REML/BLUP/GWS implementa a GWS para algumas situações.

REFERÊNCIAS

AITKEN, A. C. Studies in practical mathematics: the evaluation of the latent roots and latent vectors of a matrix. **Proceedings of the Royal Society of Edinburgh**, v. 57, p. 269-304, 1937.

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

ANDERSON, R.; BANCROFT, T. **Statistical theory in research**. New York: McGraw Hill, 1952. 399 p.

ANDRUS, D. F.; McGILLARD, L. D. Selection of dairy cattle for overall excellence. **Journal of Dairy Science,** v. 58, p. 1876-1879, 1975.

ANNICCHIARICO, P. Cultivar adaptation and recomendation from alfalfa trials in Northern Italy. **Journal of Genetics and Plant Breeding***,* v. 46, p. 269-278, 1992.

APIOLAZA, L. A.; GILMOUR, A. R.; GARRICK, D. J. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. **Canadian Journal of Forestry Research**, v. 30, p. 645-654, 2000.

ARANGO, J.; MISZTAL, I.; TSURUTA S.; CULBERTSON, M.; HERRING, W. Estimation of variance components including competitive effects of Large White growing gilts. **Journal of Animal Science**, v. 83, p. 1241-1246, 2005.

ATKINSON, A. C. The use of residuals as a concomitant variable. **Biometrika**, v. 56, p. 33-41, 1969.

BADILLA, Y.; MURILLO, O. Propuesta de un diseño de parcela para la investigación con espécies nativas en Costa Rica. **Kurú**, v. 25, p. 4-5, 1999.

BANZATO, D. A.; KRONKA, S. N. **Experimentação Agrícola***.* Jaboticabal: FUNEP, 1989. 247 p.

BARBIN, D. **Planejamento e análise estatística de experimentos agronômicos**. Arapongas: Midas, 2003. 208 p.

BARDORFF-NIELSEN, O. Plausibility inference. **Journal of the Royal Statistical Society**, **Series B**, v. 38, p. 103-131, 1976.

BARNDORFF-NIELSEN, O. E. Inference on full or partial parameters, based on the standardised log likelihood ratio. **Biometrika**, v. 73, p. 307-322, 1986.

BARNDORFF-NIELSEN, O. E. On a formula for the distribution of the maximum likelihood estimator. **Biometrika**, v. 70, p. 343-365, 1983.

BARTLETT, M. S. Nearest neighbour models in the analysis of field experiments. **Journal of the Royal Statistical Society**, **Series B**, v. 40, p. 147-174, 1978.

BARTLETT, M. S. The approximate recovery of information from replicated field experiments with large blocks. **Journal of Agricultural Science**, v. 28, p. 418-427, 1938.

BENNETT, C. A.; FRANKLIN, N. L. **Statistical analysis in chemistry and the chemical industry**. New York: J. Wiley & Sons, 1963. 724 p.

BERK, K. Computing for incomplete repeated measure. **Biometrics**, v. 43, n. 1, p. 385-398, 1987.

BERNARDO, R. Prospects for genome wide selection for quantitative traits in maize. **Crop Science**, v. 47, p.1082-1090, 2007.

BESAG, J. Spatial interaction and the statistical analysis of lattice systems. **Journal of the Royal Statistical Society**, **Series B**, v. 36, p. 192-236, 1974.

BESAG, J.; KEMPTON, R. A. Statistical analysis of field experiments using neighbouring plots. **Biometrics**, v. 42, p. 231-251, 1986.

BLASCO, A. The Bayesian controversy in animal breeding. **Journal of Animal Science**, v. 79, p. 2023-2046, 2001.

BOX, G. E. P.; COX, D. R. An analysis of transformation. **Journal of the Royal Statistics Society Series B**, v. 26, p. 211-243, 1964.

BOX, G. E. P.; TIAO, G.C. **Bayesian inference in statistical analysis**. Reading: Addison-Wesley, 1973. 588 p. (Addison-Wesley Series in Behavioral Science: Quantitative Methods).

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v. 88, p. 9-25, 1993.

BRIAN, C. Interprétation statistique des essais deux voies: décomposition factorielle des résidus et étude de la structure des interactions. **Annalles a'Amelioration des Plantes**, v. 28, p. 395-409, 1978.

BROWN, H.; KEMPTON, R. A. The application of REML in clinical trials. **Statistics in Medicine**, v. 16, p. 1601-1617, 1994.

BROWN, H.; PRESCOTT, R. **Applied mixed models in medicine**. Chichester: J. Willey & Sons, 1999. 408 p.

BUENO FILHO, J. S. S. **Uso de modelos mistos na predição de valores genéticos aditivos em testes de progênies florestais**. 118 f. 1997. Tese (Doutorado) – ESALQ, Piracicaba.

BUENO FILHO, J. S. S.; GILMOUR, S. Planning incomplete block experiments when treatments are genetically related. **Biometrics**, v. 59, p. 375-381, 2003.

CALUS, M. P. L.; MEUWISSEN, T. H. E.; ROOS, A. P. W.; VEERKAMP, R. F. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 178, p. 553-561, 2008.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of breeding value when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. **Journal of Animal Breeding and Genetics**, v. 124, p. 362-368, 2007.

CAMPOS, G.; GIANOLA, D. Factor analysis models for structuring covariance matrices of additive genetic effects: a bayesian implementation. **Genetics Selection Evolution**, v. 39, n. 5, p. 491-494, 2007.

CAMPOS, H. **Estatística aplicada à experimentação com cana-de-açúcar**. Piracicaba: FEALQ, 1984. 292 p.

CAPPA, E. P.; CANTET, R. J. C. Bayesian estimation of direct and competition additive covariance in individual mixed models. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte, Intituto Prociência, 2006. 1 CD-ROM.

CARBONELL, S. A. M.; CHIORATO, A.; RESENDE, M. D. V. de; DIAS, L. A. S.; BERALDO, A. L. A.; PERINA, E. F. Estabilidade de cultivares e linhagens de feijoeiro em diferentes ambientes no Estado de São Paulo. **Bragantia**, v. 66, n. 2, p. 193-201, 2007.

CARLIN, B. P.; LOUIS, T. A. **Bayes and empirical Bayes methods for data analysis**. London: Chapman and Hall, 1996. 399 p.

CAVALCANTI, J. J. V.; RESENDE, M. D. V. de; CRISÓSTOMO, J. R.; BARROS, L. M.; PAIVA, J. R. de. Genetic control of quantitative traits and hybrid breeding strategies for cashew improvement. **Crop Breeding and Applied Biotechnology**, v. 7, p. 186-195, 2007.

CEAPOIU, N. **Metode statistice aplicate in experientele agricole si biologice**. Bucuresti: Editura Agro-Silvica, 1968. 550 p.

CHALONER, K.; VERDINELLI, I. Bayesian experimental design: a review. **Statistical Science**, v. 10, n. 3, p. 237-304, 1995.

CHEW, V. Multiple comparison procedures: uses, abuses and alternatives. In: WORKSHOP OF THE GENETICS AND BREEDING OF SOUTHERN FOREST TREES, 1986, Gainesville. **Proceedings**... Gainesville: University of Florida, 1986. p. 48-58.

CHOO, T. M.; REINBERGS, E. Analyses of skewness and kurtosis for detecting gene interaction in a doubled-haploid population. **Crop Science**, v. 22, p. 231-234, 1982.

COCHRAN, W. G. Improvement by means of selection. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 2., 1951, Berkeley. **Proceedings**... Berkeley: University of California Press, 1951. p. 449-470.

COCHRAN, W. G.; COX, G. M. **Experimental designs**. 2. ed. New York: J. Wiley, 1957. 611 p.

COMREY, A. L.; HOWARD, B. L. **A first course in factor analysis**. Mahwah: L. Erlbaum Associates , 1992. 448 p.

COOPER, D. M.; THOMPSON, R. A note on the estimation of the parameters of the autoregressive moving average process. **Biometrika**, v. 64, p. 625-628, 1977.

CORNELIUS, P. L.; CROSSA, J. Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. **Crop Science**, v. 39, p. 998-1009, 1999.

CORNELIUS, P. L.; CROSSA, J.; SEYEDSADR, M. S. Statistical tests and estimators of multiplicative models for genotype by environment interaction. In: KANG, M. S.; GAUCH, H. G. **Genotype by environment interaction**. Boca Raton: CRC Press, 1996. p. 199- 234.

CORNELIUS, P. L.; SEYEDSADR, M. S.; CROSSA, J. Using the shifted multiplicative model to search for separability in crop cultivar trials. **Theoretical and Applied Genetics**, v. 84, p. 161-172, 1992.

CORREL, R. L.; ANDERSON, R. B. Removal of intervarietal competition effects in forestry varietal trials. **Silvae Genetica**, v. 32, n. 5-6, p. 162-165, 1983.

COSTA e SILVA, J.; DUTKOWSKI, G. W.; BORRALHO, N. M. N. Across-site heterogeneity of genetic and environmental variances in the genetic evaluation of Eucalyptus globules trials for height growth. **Annals of Forest Science**, v. 62, p.183-191, 2005.

COSTA e SILVA, J.; DUTKOWSKI, G. W.; GILMOUR, A. R. Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. **Silvae Genetica**, v. 31, p. 1887-1893, 2001.

COTERILL, P. P.; JAMES, J. Number of offspring and plot sizes required for progeny testing. **Silvae Genetica**, Frankfurt, v. 23, n. 6, p. 203-208, 1984.

COX, D. R.; REID, N. Parameter orthogonality and approximate conditional inference. **Journal of the Royal Statistical Society***,* **Series B**, v. 49, p. 1-39, 1987.

CRESSIE, N. A. C. **Statistics for spatial data analysis**. New York: J. Wiley, 1993. 900 p.

CRIANICEANU, C. M.; RUPPERT, D. Likelihood ratio tests in linear mixed models with one variance component. **Journal of the Royal Statistical Society**, **Series B**, v. 66, p. 165-185, 2004.

CROSSA, J. Statistical analysis of multi-location trials. **Advances in Agronomy**, v. 44, p. 55-85, 1990.

CROSSA, J.; GAUCH, H. G.; ZOBEL, R. W. Additive main effects and multiplicative interaction analysis of two international maize cultivars trials. **Crop Science**, v. 30, n. 3, p. 493-500, 1990.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa, MG: Universidade Federal de Viçosa, 2003. v. 2. 585 p.

CRUZ, C. D.; REGAZZI, O. J. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa, MG: Universidade Federal de Viçosa: Imprensa Universitária, 1994. 390 p.

CULLIS, B. R. Aerobic geostatistical mixed model – rotating and stretching. In: ANALYSING GROUPED DATA WORKSHOP, 2005, Canberra. **Proceedings**... Canberra: Australian National University, 2005. Disponível em: http://www.statsoc.org.au/Branches/Canberra/Workshop%20files/BCullis.pdf. Acesso em: 15 jan. 2007.

CULLIS, B. R.; GLEESON, A. C. Spatial analysis of field experiments-an extension at two dimensions. **Biometrics**, v. 47, p. 1449-1460, 1991.

CULLIS, B. R.; GOGELL, B.; VERBYLA, A.; THOMPSON, R. Spatial analysis of multi-environment early generation variety trials. **Biometrics**, v. 54, p. 1-18, 1998.

CULLIS, B.; SMITH, A. B.; COOMBES, N. On the design of early generation trials with correlated data. **Journal of Agricultural, Biological and Environmental Statistics**, v. 11, n. 4, p. 381-393, 2006.

CULLIS, B.; SMITH, A. B.; THOMPSON, R. Perspectives of ANOVA, REML and a general linear mixed model. In: ADAMS, N. M.; CROWDER, M. J.; HAND D. J.; STEPHENS D. A. (Ed.). **Methods and models in statistics in honour of Professor John Nelder FRS**. London: University of London, London College, 2004. p. 53-94.

CULLIS, B.; SMITH, A. B.; VERBYLA, A. P.; WELHAM, S. J.; THOMPSON, R. **Mixed models for data analysts**. New York: CRC Press, 2007. 288 p.

DAETWYLER, H. D.; VILLANUEVA, B.; BIJMA, P.; WOOLIAMS, J. A. Inbreeding in genome-wide selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 369-376, 2007.

DAVIDIAN, M.; GILTINAN, D. M. Some general estimation methods for non-linear mixed effects models. **Journal of Biopharmaceutics Statistics**, v. 3, p. 23-55, 1993.

DAVIS, J. C. **Statistics and data analysis in geology**. New York: J. Wiley, 1986. 646 p.

DEKKERS, J. C. M. Commercial application of marker and gene assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v. 82, p.313-328, 2004.

DEKKERS, J. C. M. Prediction of response to marker assisted and genomic selection using selection index theory. **Journal of Animal Breeding and Genetics**, v. 124, p. 331-341, 2007.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistic Society**, London, v. 39, p. 1-38, 1977.

DIAS, L. A. S. dos. **Melhoramento genético do cacaueiro**. Viçosa, MG: FUNAPE, 2001. 578 p.

DIAS, L. A. S. dos; RESENDE, M. D. V. de. Estratégias e métodos de seleção. In: DIAS, L. A. S. dos. (Org.). **Melhoramento genético do cacaueiro**. Viçosa, MG: FUNAPE, 2001b. p. 217-287.

DIAS, L. A. S. dos; RESENDE, M. D. V. de. Experimentação no melhoramento. In: DIAS, L. A. S. dos. (Org.). **Melhoramento genético do cacaueiro**. Viçosa, MG: FUNAPE, 2001a. p. 439-492.

DIAS. L. A. S. dos. Análises multidimensionais. In: ALFENAS, A. C. (Ed.). **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microorganismos. Viçosa, MG: Ed. da Universidade Federal de Viçosa, 1998. p.

DICKERSON, G. E. Implications of genotype – environmental interactions in animal breeding. **Animal Production**, v. 4, p. 47-63, 1962.

DOBSON, A. J. **An introduction to generalized linear models**. Melbourne: Chapman & Hall, 1990. 174 p.

DRAPER, N. R.; GUTTMAN, I. Incorporating overlap effects from neighbouring units into response surface models. **Applied Statistics**, v. 9, p. 128-134, 1980.

DUARTE, J. B. **Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal**. 2000. 293 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – ESALQ, Piracicaba.

DUARTE, J. B.; VENCOVSKY, R. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agrícola**, v. 58, n. 1, p. 109-117, 2001.

DUARTE, J. B; VENCOVSKY, R. **Interação genótipos x ambientes: uma introdução a análise AMMI**. Ribeirão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Serie Monografias, n. 9)

DUDEWICZ, E. J. Introduction to statistics and probability. In: DUDEWICZ, E. J**. Ranking and selection procedures**. New York: Holt: Rinehart & Winston, 1976. p. 315-383.

DURBAN, M.; CURRIE, I. D. Adjustment of the profile likelihood for a class of normal regression models. **Scandinavian Journal of Statistics**, v. 27, p. 535-542, 2000.

DURBAN, M.; CURRIE, I.; KEMPTON, R. Adjusting for fertility and competition in variety trials. **Journal of Agricultural Science**, v. 136, p. 129-149, 2001.

DURBAN, M.; HACKET, C.; CURRIE, I. Blocks, trends and interference in field trials. In: INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING, 14., Graz. **Proceedings**... Graz: Statistical Modelling Society, 1999. p. 492-495.

DUTKOWSKI, G. W.; McRAE, T. A.; POWELL, M. B.; PILBEAM, D. J.; JOYCE, K.; TIER, B.; KERR, R. Benefits from data and pedigree integration in genetic evaluation. In: AUSTRALASIAN PLANT BREEDING CONFERENCE, 13., 2006, Christchurch. **Proceedings***.* Christchurch: Australasian Plant Breeding Association, 2006. 1 CD-ROM.

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties**. Crop Science**, v. 6, p. 36-40, 1966.

EDWARDS, A. W. F. **Likelihood**. Cambridge: Cambridge University Press, 1972. 235 p.

EEUWIJK, F. A. van; KEIZER, L. C. P.; BAKKER, J. J. Linear and bilinear models for the analysis of multi-environment trials. II. An application to data from Dutch Maize Variety Trials. **Euphytica**, v. 84, p. 9-22, 1995.

EFRON, B. 1975. Biased versus unbiased estimation. **Advances in Mathematics**, v. 16, p. 259-277, 1975.

EFRON, B. Why isn't everyone a bayesian? **American Statistician**, v. 40, p. 11, 1986.

EFRON, B.; MORRIS, C. Data analysis using Stein's estimator and its generalizations. **Journal of the American Statistical Association**, v. 70, n. 350, p. 311-319, 1975.

EFRON, B.; MORRIS, C. Stein's paradox in statistics. **Scientific American**, v. 236, n. 5, p. 119-127, 1977.

ELSTON, D. A. Estimating of denominator degrees of freedom of F-distributions for assessing Wald statistics for fixed-effect factors in unbalanced mixed models. **Biometrics**, v. 54, n. 3, p. 1085-1096, 1998.

ENGEL, B.; KEEN, A. A simple approach for the analysis of generalized linear mixed models. **Statistica Neerlandica**, v. 48, n. 1, p. 1-22, 1994.

ENGEL, B.; KEEN, A. An introduction to generalized linear mixed models. In: INTERNATIONAL BIOMETRIC CONFERENCE: Invited Papers, 18., 1996, Amsterdam. **Proceedings**… Amsterdam: International Biometric Society, 1996. p. 125-135.

FEDERER, W. T. Augmented designs. **Biometrics**, v. 14, p. 134, 1958.

FEDERER, W. T. **Experimental designs**: theory and application. New Delhi: Oxford Publ., 1955. 544 p.

FEDERIZZI, L. C.; MILACH, S. C. K.; PACHECO, M. T. Melhoramento da Aveia. In: BORÉM, A. (Ed.). **Melhoramento de espécies cultivadas**. Viçosa, MG: Ed. da Universidade Federal de Viçosa, 1999. p. 131-157.

FERNANDEZ, G. C. J. Residual analysis and data transformations: important tools in statistical analysis. **Hortscience**, v. 27, n. 4, p. 297-300, 1992.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant breeding programme. **Australian Journal of Agricultural Research**, v. 14, p. 742-754, 1963.

FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London, Series A**, n. 222, p. 309-368, 1922.

FISHER, R. A. **Statistical methods for research workers**. 10. ed. New York: Hafner, 1948. 345 p.

FISHER, R. A. **Statistical methods for research workers**. London: Oliver and Boyd, 1925. 314 p.

FISHER, R. A. The arrangement of field experiments. **Journal of the Ministry of Agriculture of Great Britain**, v. 33, p. 503-513, 1926.

FISHER, R. A.; IMMER, F. R.; OLOF, T. The genetical interpretation of statistics of third degree in the study of the quantitative inheritance. **Genetics**, v. 17, n. 2, p 107-124, 1932.

FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation. II. The manurial response of different potato varieties. **Journal of Agricultural Science**, v. 13, p. 311-320, 1923.

FOULLEY, J. L. Algorithme EM: Théorie et application au modèle mixte. **Journal de la Société Française de Statistique**, v. 143, p. 57-109, 2002.

FOULLEY, J. L. **Le modéle linéaire mixte**. Paris: INRA, 2003. 139 p.

FOULLEY, J. L.; DELMAS, C.; ROBERT-GRANIER, C. Méthods du maximum de vraisemblance in modéle linéaire mixte. **Journal de la Société Française de Statistique**, v. 143, p. 5-52, 2002.

FOULLEY, J. L.; DYK, D. A. van. The PX-EM algorithm for fast and stable fitting of Henderson's mixed model. **Genetics, Selection, Evolution**, v. 32, p. 143-163, 2000.

FOULLEY, J. L.; QUAAS, R. L. Heterogeneous variances in gaussian linear mixed models. **Genetics Selection Evolution**, v. 27, p. 211–228, 1995.

FU, Y.; CLARKE, G. P. Y.; NAMKOONG, G.; YANCHUK, A. D. Incomplete block designs for genetic testing: statistical efficiencies of estimating family means. **Canadian Journal of Forest Research**, v. 28, n. 7, p. 977-986, 1998.

GALLAIS, A. Sur quelques aspects de la competition en amelioration des plantes. **Annales des Ameliorations des Plantes**, v. 25, p. 51-64, 1975.

GALLAIS, A. **Théorie de la sélection en amélioration des plantes**. Paris: Mason, 1989. 588 p.

GALLO, J.; KHURI, A. I. Exact tests for the random and fixed effects in an unbalanced mixed-two-way cross-classification model. **Biometrics**, v. 46, n. 3, p. 1087-1095, 1990.

GAMERMAN, D. **Simulação estocástica via cadeias de Markov**. Caxambu: Associação Brasileira de Estatística, 1996. 196 p.

GARZA, A. M. **Diseño y análisis de experimentos con caña de azúcar**. Chapingo: Universidad de Chapingo, 1972. 204 p.

GAUCH, H. G. Model selection and validation for yield trials with interaction. **Biometrics**, v. 44, p. 705-715, 1988.

GAUCH, H.G. **Statistical analysis of regional yield trials: AMMI analysis of factorial designs**. Amsterdam: Elsevier, 1992. 172 p.

GEORGE, A.; LIU, J. **Computer solution of large sparse positive definite systems**. Englewood Cliffs: Prentice Hall, 1981. 324 p.

GIANOLA, D. Advances in Bayesian methods for quantitative genetic analysis. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**. Paris: INRA, 2002. p. 215-222.

GIANOLA, D. Can BLUP and REML be improved upon? In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 4., 1990. Joyce Darling. **Proceedings**... Edinburgh: University of Edinburgh, 1990. v. 13, p. 445-449.

GIANOLA, D. Statistics in animal breeding. In: RAFFERTY, A. E.; TANNER, M. A.; WELLS, M. T. **Statistics in the 21st century**. London: Chapman & Hall, 2001. 535 p.

GIANOLA, D. Statistics in animal breeding: angels and demons. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte, Intituto Prociência, 2006. 1 CD-ROM.

GIANOLA, D.; FERNANDO, R. L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v. 63, p. 217-244, 1986.

GIANOLA, D.; FERNANDO, R. L.; IM, S.; FOULLEY, J. L. Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems. **Genome**, Ottawa, v. 31, p. 768-777, 1989.

GIANOLA, D.; HAMMOND, K**. Advances in statistical methods for genetic improvement of livestock**. Berlin: Springer-Verlag, 1990. 534 p.

GILMOUR, A. R. Mixed model regression mapping for QTL detection in experimental crosses. **Computational Statistics and Data Analysis**, v. 51, n. 8, p. 3749-3764, 2007.

GILMOUR, A. R. Post blocking gone too far! Recovery of information and spatial analysis in field experiments. **Biometrics***,* v. 56, p. 944-946, 2000.

GILMOUR, A. R. Statistical models for multidimensional (longitudinal/spatial) data. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte, Instituto Prociência, 2006. 1 CD-ROM.

GILMOUR, A. R.; CULLIS, B. R.; VERBYLA, A. P. Accounting for natural and extraneous variation in the analysis of field experiments. **Journal of Agricultural, Biological and Environmental Statistics**, v. 2, p. 269-293, 1997.

GILMOUR, A. R.; CULLIS, B. R.; WELHAM, S. J.; GOGEL, B. J.; THOMPSON, R. An efficient computing strategy for prediction in mixed linear models. **Computational Statistics and Data Analysis**, v. 44, n. 4, p. 571-586, 2004.

GILMOUR, A. R.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. **ASReml Reference manual**: release 1.0. 2. ed. Harpenden: Rothamsted Research, Biomathematics and Statistics Department, 2002. 187 p.

GILMOUR, A. R.; CULLIS, B. R; FRENSHAM, A. B.; THOMPSON, R. (Co)variance structures for linear models in the analysis of plant improvement data. In: SYMPOSIUM ON COMPUTATIONAL STATISTICS, 3., 1998, Bristol. **Proceedings in computational statistics**. Heidelberg: Physica-Verlag, 1998. p. 53-64. COMPSTAT'98.

GILMOUR, A. R.; THOMPSON, R. Equation ordering for average information REML. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte, Instituto Prociência, 2006. 1 CD-ROM.

GILMOUR, A. R.; THOMPSON, R. Estimating parameters of a singular variance matrix in ASREML. In: AUSTRALASIAN GENSTAT CONFERENCE**,** 2002, Busselton. **Proceedings**... Busselton: Atlas Conferences Inc. 2002.

GILMOUR, A. R.; THOMPSON, R. Modelling variance parameters in ASREML for repeated measures. In: WORLD CONGRESS ON GENETIC APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. **Proceedings**… Armidale: AGBU: University of New England, 1998. v. 27, p. 453-454.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average information REML: an efficient algorithm for parameter estimation in linear mixed models. **Biometrics**, v. 51, p. 1440-1450, 1995.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R.; WELHAM, S. J. ASREML estimates variance matrices from multivariate data using the animal model. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**... Paris: INRA, 2002. 1 CD-ROM. (Communication, n. 28).

GIURGIU, V. **Aplicatii ale statisticii matematice in silvicultura**. Bucuresti: Editura Agro-Silvica, 1966.

GLEESON, A. C. Spatial analysis. In: KEMPTON, R. A.; FOX, P. N. **Statistical methods for plant variety evaluation**. London: Chapman & Hall, 1997. p. 68-85.

GLEESON, A. C.; CULLIS, B. R. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. **Biometrics**, v. 43, p. 277-288, 1987.

GLENDINNING, D. R.; VERNON, A.J. Inter-varietal competition in cocoa trials. **Journal of Horticultural Science**, v. 40, p. 317-319, 1965.

GODDARD, M. E. A mixed model for analysis of data on multiple genetic markers**. Theoretical and Applied Genetics**, v. 83, p. 878-886, 1992.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.

GOLLOB, H. F. A statistical model which combines features of factor analytic and analysis of variance technique. **Psychometrika**, v. 33, n. 1, p. 73-115, 1968.

GOMES, M. I. Verossimilhança e inferencia estatística. In: COLÓQUIO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL, 2., 1981. **Actas**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 1981. p. 230-246.

GRASER, H. U.; SMITH, S. P.; TIER, B. A derivative free approach for estimating variance components in animal models by restricted maximum likelihood. **Journal of Animal Science**, Champaign, v. 64, n. 5, p. 1362-1370, 1987.

GREEN, P. J.; JENNISON, C.; SEHEULT, A. Analysis of field experiments by least square smoothing. **Journal of the Royal Statistical Society, Series B**, v. 47, n. 2, p. 299-315, 1985.

GREEN, P. J.; SILVERMAN, B. W. **Nonparametric regression and generalized linear models**. London: Chapman & Hall, 1994. 182 p.

GRONDONA, M. O.; CRESSIE, N. Using spatial considerations in the analysis of experiments. **Technometrics**, v. 33, p. 381-392, 1991.

GRONDONA, M. O.; CROSSA, J.; FOX, P. N.; PFEIFFER, W. H. Analysis of variety yield trials using two-dimensional separable ARIMA processes. **Biometrics**, v. 52, p. 763-770, 1996.

GRUBER, M. H. J. **Improving efficiency by shrinkage. The James-Stein and ridge regression estimators**. New York: Marcel Dekker, 1998. 632 p.

HAAPANEN, M. Effect of plot size and shape on the efficiency of progeny tests. **Silva Fennica**, v. 26, n. 4, p. 201-209, 1992.

HARTLEY, H. O.; RAO, J. N. K. Maximum likelihood estimation for the mixed analysis of variance model. **Biometrika**, v. 54, p. 93-108, 1967.

HARVILLE, D. A. Making REML computationally feasible for large data sets: use of the Gibbs sampler. **Journal of Statistical Computation and Simulation**, v. 74, p. 135-154, 2004.

HARVILLE, D. A. **Matrix algebra from a statistician perspective**. New York: Springer Verlag, 1997. 630 p.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American Statistical Association**, v. 72, n. 2, p. 320-328, 1977.

HARVILLE, D. A.; CARRIQUIRY, A. L. Classical and Bayesian prediction as applied to unbalanced mixed linear models. **Biometrics**, v. 48, p. 987-1003, 1992.

HAWKINS, D. M. **Identification of outliers**. London: Chapman & Hall, 1980. 180 p.

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

HEGYI, F. A simulation model for managing jack pine stands. In: FRIES, J. (Ed.). **Growth models for tree and stand simulation**. Stockholm: Department of Forest Resources, 1974. p. 74-85.

HENDERSON JUNIOR, C. R.; HENDERSON, C. R. Analysis of covariance in mixed models with unequal subclass numbers. **Communications in Statistics**, v. A8, p. 751, 1979.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, v. 32, p. 69-83, 1976.

HENDERSON, C. R. **Aplications of linear models in animal breeding**. Guelph: University of Guelph, 1984. 462 p.

HENDERSON, C. R. Best linear estimation and prediction under a selection model. **Biometrics**, v. 31, p. 423-447, 1975.

HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, v. 9, p. 226-252, 1953.

HENDERSON, C. R. Estimation of variances in animal model and reduced animal model for single traits and single records. **Journal of Dairy Science**, v. 69, p. 1394-1402, 1986.

HENDERSON, C. R. **Sire evaluation and genetic trends**. In: ANIMAL BREEDING AND GENETICS SYMPOSIUM IN HONOUR OF J. LUSH, 1973, Champaign. **Proceedings.**.. Champaign: American Society of Animal Science, 1973. p.10-41.

HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relatives records. **Journal of Animal Science**, v. 3, p. 1188-1197, 1976.

HICKS, C. R. **Fundamental concepts in the design of experiments**. New York: Holt, 1973. 349 p.

HILL, R. R.; ROSENBERGER, J. L. Methods for combining data from germplasm evaluation trials. **Crop Science**, v. 25, p. 467-470, 1985.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Análise exploratória de dados**: técnicas robustas, um guia. Lisboa: Salamandra, 1983. 446 p.

HOFER, A. Variance component estimation in animal breeding: a review. **Journal of Animal Breeding and Genetics**, v. 115, p. 247-265, 1998.

HOULE, D. Comparing evolvability and variability of quantitative traits. **Genetics**, v. 130, p.195-204, 1992.

IEMMA, A. F. **Modelos lineares**: uma introdução para profissionais da pesquisa agropecuária. Londrina: UEL: RBRAS, 1987. 262 p.

ISAACKS, E.; SRIVASTAVA, R.M. **An introduction of Applied Geostatistics**. New York: Oxford University Press, 1989. 561 p.

JAFFREZIC, F.; MEZA, C.; LAVIELLE, M.; FOULLEY, J. L. Genetic analysis of growth curves using the SAEM algorithm. **Genetics Selection Evolution**, v. 38, n. 6, p. 583-600, 2007.

JAFFREZIC, F.; PLETCHER, S. D. Statistical models for estimating the genetic basis of repeated measures and other function valued traits. **Genetics**, v. 156, p. 913-922, 2000.

JAFFREZIC, F.; THOMPSON, R.; HILL, W. G. Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. **Genetical Research**, v. 82, p. 55-65, 2003.

JAFFREZIC, F.; WHITE, I. M. S.; THOMPSON, R. Use of the score test as a goodness-of-fit measure of the covariance structure in genetic analysis of longitudinal data. **Genetics Selection Evolution**, v. 35, n. 2, p. 185-198, 2003.

JAFFREZIC, F.; WHITE, I. M. S.; THOMPSON, R.; VISSCHER, P. M; HILL, W. G. Statistical models for the genetic analysis of longitudinal data. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**. Paris: INRA, 2002. p. 1-4. 1 CD-ROM. (Communication, n. 16-08).

JAMES, W.; STEIN, C. Estimation with quadratic loss. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 4., 1961, Berkeley: **Proceedings**... Berkeley: University of Berkeley, 1961. p. 361-379.

JENNRICH, R. L.; SCHLUCHTER, M. D. Unbalanced repeated measures models with structured covariance matrices. **Biometrics**, v. 42, p. 805-820, 1986.

JENSEN, J.; MANTYSAARI, E. A.; MADSEN, P.; THOMPSON, R. Residual maximum likelihood estimation of covariance components in multivariate mixed linear models using average information. **Journal of the Indian Society of Agricultural Statistics**, v. 49, p. 215-236, 1997.

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of Dairy Science**, v. 78, p. 449-456, 1995.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. Englewood : Prentice Hall Inc., 1988. 594 p.

JOYCE, D.; FORD, R.; FU, Y. B. Spatial patterns of tree height variations in a black spruce farm-field progeny test and neighbors-adjusted estimations of genetic parameters. **Silvae Genetica**, v. 51, n. 1, p. 13-18, 2002.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v. 23, p. 187-200, 1958.

KEEN, A.; ENGEL, B. Analysis of a mixed model for ordinal data by iterative re-weighted REML. **Statistica Neerlandica**, v. 51, n. 2, p. 129-144, 1997.

KEMPTHORNE, O. **Design and analysis of experiments**. New York: J. Wiley, 1952. 631 p.

KEMPTON, R. A. Adjustment for competition between varieties in plant breeding trials. **Journal of Agricultural Science**, Cambridge, v. 98, p. 599-611, 1982.

KEMPTON, R. A. Discussion on The analyses of designed experiments and longitudinal data using smootthing splines. **Journal of the Royal Statistical Society, Series C**, v. 48, p. 300-301, 1999.

KEMPTON, R. A. Statistical models for interplot competition. **Aspects of Applied Biology**, v. 10, p. 11-120, 1985.

KEMPTON, R. A. The use of biplots in interpreting variety by environment interactions. **Journal of Agricultural Science**, v. 103, p. 123-135, 1984.

KEMPTON, R. A.; HOWES, C. W. The use of neighbouring plot values in the analysis of variety trials. **Applied Statistics**, v. 30, p. 59-70, 1981.

KEMPTON, R. A.; SERAPHIN, J. C.; SWORD, A. M. Statistical analysis of two dimensional variation in variety yield trials. **Journal of Agricultural Science**, v. 122, p. 335-342. 1994.

KENDALL, M. G. Factor Analysis. **Journal of the Royal Statistical Society, Series B**, v. 12, p. 60-94, 1950.

KENNEDY, B. W. Charles Roy Henderson: The unfinished legacy. **Journal of Dairy Science**, v. 74, n. 11, p. 4067-4081, 1991.

KENNEDY, B. W.; TRUS, D. Considerations on genetic connectedeness between management units under an animal model. **Journal of Animal Science**, v. 71, p. 2341-2352, 1993.

KENWARD, M. G.; ROGER, J. H. Small sample inference for fixed effects from restricted maximum likelihood. **Biometrics**, v. 53, n. 3, p. 983-997, 1997.

KEULS, M. The use of the studentized range in connection with an analysis of variance. **Euphytica**, v. 1, p. 112-122, 1952.

KHURI, A. I.; LITTELL, R. C. Exact tests for the main effects variance components in an unbalanced random two-way model. **Biometrics**, v. 43, n. 2, p. 545-560, 1987.

KIRKPATRICK, M.; HILL, W. G.; THOMPSON, R. Estimating the covariance structure of traits during growth and ageing, illustrated with lactations in dairy cattle. **Genetical Research**, v. 64, p. 57-69, 1994.

KOLBEHDARI, D.; SHAEFFER, L. R.; ROBINSON, J. A. B. Estimation of genome-wide haplotype effects in half-sib designs. **Journal of Animal Breeding and Genetics**, v. 124, p. 356-361, 2007.

KONSTANTINOV, P. N.; PLOTNIKOV, N. I. **Cu privire la metodica experientelor in cimp**. Bucuresti: Ed. Analele Romano - Sovietice Agricultura - Zootehnie, 1960.

KOUTSOYIANIS, A. **Theory of econometrics**. London: Mac Millan Press, 1973. 631 p.

KUSNANDAR, D. **The identification and interpretation of genetic variation in forestry plantation**. Perth: The University of Western Australia, 2001. 3 p. PhD Thesis Abstract.

LANDE, R.; THOMPSON, R. Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics, v. 124, p.743-756, 1990.

LANGE, K. **Mathematical and statistical methods for genetic analysis**. New York: Springer-Verlag, 1998. 290 p.

LAVIELLE, M.; MEZA, C. A parameter expansion version of the SAEM algorithm. **Statistics and Computing**, v. 17, n. 2, p. 121-130, 2007.

LAWLEY, D. N; MAXWELL, A. E. **Factor analysis as a statistical tool**. 2nd ed. London: Lawrence Erlbaum Associates, 1992. 448 p.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society, Series B**, v. 58, p. 619-678, 1996.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models: a synthesis of generalized linear models, random effects model and structured dispersions. **Biometrika**, v. 88, p. 987-1006, 2001.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized linear models with random effects: unified analysis via H – likelihood**. London: Chapman & Hall, 2007. 416 p.

LEGARRA, A.; MISZTAL, I. Computing strategies in genome-wide selection. **Journal of Dairy Science**, v. 91, n.1, p. 360-366, 2008.

LEHMANN, E. L. Some model I problems of selection. **Annals of Mathematical Statistics**, v. 32, p. 990, 1961.

LEONARDECZ-NETO, E. **Competição intergenotípica na análise de testes de progênies em essências florestais**. 60 f. Tese (Doutorado) – ESALQ, Piracicaba.

LIANG, K. Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, v. 73, p. 13-22, 1986.

LIN, C. S.; BINNS, M. R. A superiority measure of cultivar performance for cultivar x location data. **Canadian Journal of Plant Science**, Ottawa, v. 68, n. 3, p. 193-198, 1988.

LINDLEY, D. V. The bayesian approach. **Scandinavian Journal of Statistics**, v. 5, p. 1-26, 1978.

LINDSEY, J. K. Some statistical heresies. **Journal of the Royal Statistical Society, Series B,** v. 48, p. 1-40, 1999.

LITTELL, R. C. Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. **Journal of Agricultural, Biological and Environmental Statistics**, v. 7, n. 4, p. 472-491, 2002.

LIU, C.; RUBI, D. B.; WU, Y. N. Parameter expansion to accelerate EM: the PX-EM algorithm. **Biometrika**, v. 85, p. 755-770, 1998.

LONG, N.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A; AVENDAÑO, S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, v.124, p. 377–389, 2007.

LOO-DINKS, J. A.; TAUER, C. G. Statistical efficiency of six progeny test field designs on three loblolly pine (*Pinus taeda* L.) site types. **Canadian Journal of Forest Research**, v. 17, p. 1066-1070, 1987.

LOTODÉ, R.; LACHENAUD, P. Méthodologie destinée aux essais de sélection du cacaoyer. **Café Cacao Thé***,* v. 32, n. 4, p. 275-292, 1988.

LUSH, J. L. Family merit and individual merit as bases for selection. **American Naturalist**, v. 81, p. 241-261, 1947.

LUSH, J. L. The number of daughters necessary to prove a sire. **Journal of Dairy Science**, v. 14, p. 209-220, 1931.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, Inc., 1997. 980 p.

MACHADO, A. de A.; SILVA, J. G. C.; DEMÉTRIO, C. G. B.; FERREIRA, D. F. Estatística experimental: uma abordagem baseada no planejamento e no uso de recursos computacionais. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 50.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 11., 2005, Londrina. **Reunião**... [S.l.]: The International Biometric Society, 2005. p.30-78. RBRAS.

MAGNUSSEN, S. A method to adjust simultaneously for spatial microsite and competition effects. **Canadian Journal of Forestry Research**, v. 24, p. 985-995, 1994.

MAGNUSSEN, S.; YEATMAN, C. W. Adjusting for inter-row competition in a jack pine provenance trial. **Silvae Genetica**, v. 36, n. 5-6, p. 206-214, 1987.

MAHALANOBIS, P. C. On the generalized distance in statistics. **Proceedings of the National Institute of Science**, v. 2, p. 49-55, 1936.

MANDEL, J. A new analysis of variance model for non-additive data. **Technometrics**, v. 13, n. 1, p. 1-18, 1971.

MANDEL, J. Non-additivity in two-way analysis of variance. **Journal of the American Statistical Association**, v. 56, p. 878-888, 1961.

MANTYSAARI, E. A; VLECK, L. D. van. Restricted maximum likelihood estimates of variance components from multi-trait sire models with large number of fixed effects. **Journal of Animal Breeding and Genetics**, v. 106, p. 409-422, 1989.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Acadenic Press, 1988. 512 p.

MARTIN, R. J. The use of time-series models and methods in the analysis of agricultural field trials. **Communications on Statistical Theory and Methods**, v. 19, n. 1, p. 55-81, 1990.

Mc CULLAGH, P.; NELDER, J. A. **Generalized linear models**. 2. ed. London: Chapman and Hall, 1989. 511 p.

Mc CULLAGH, P.; TIBSHIRANI, R. A simple method for the adjustment of profile likelihoods. **Journal of the Royal Statistics Society, Series B**, v. 52, p. 325-344, 1990.

McCULLOCH, C. E; SEARLE, S. R. **Generalized, linear, and mixed models**. New York: J. Wiley, 2001. 325 p.

McRAE, T. A.; DUTKOWSKI, G. W.; PILBEAM, D. J.; POWELL, M. B.; TIER, B. Genetic evaluation using TREEPLAN system. In: McKEAND, S.; LI, B. **Forest genetics and tree breeding in the age of genomics**. Charlestone: North Carolina State University. 2004.

MEAD, R. A mathematical model for the estimation of interplant competition. **Biometrics**, v. 23, p. 189-205, 1967.

MEAD, R. Design of plant breeding trials. In: KEMPTON, R. A.; FOX, P. N. (Ed.). **Statistical methods for plant variety evaluation***.* London: Chapman & Hall, 1997. p. 40-67.

MENG, X. L.; DYK, D. A. van. Fast EM-type implementation for mixed effects models. **Journal of the Royal Statistics Society, Series B**, v. 60, p. 559-578, 1998.

MEUWISSEN, T. H. E. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, v. 124, p. 321-322, 2007.

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

MEYER, K. An average information restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. **Genetics, Selection, Evolution**, v. 29, p. 97-116, 1997.

MEYER, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. **Genetique, Selection, Evolution**, v. 23, p. 67-83, 1991.

MEYER, K. Estimation of genetic parameters. In: HILL, W.G.; MACKAY, T. F. C**. Evolution and animal breeding**. Wallingford: CAB International, 1989a. p. 161-167.

MEYER, K. Random regression analyses using B-splines to model growth of Australian Angus cattle. **Genetics, Selection, Evolution**, v. 37, p. 473-500, 2005.

MEYER, K. WOMBAT – digging deep for quantitative genetic analysis by restricted maximum likelihood. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

MEYER, K.; HILL, W. G. Estimation of genetic and phenotypic covariance functions for longitudinal or repeated records by restricted maximum likelihood. **Livestock Production Science**, v. 47, p. 185-200, 1997.

MEYER, K.; KIRKPATRICK, M. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. **Genetics, Selection, Evolution**, v. 37, p. 1-30, 2005.

MISZTAL, I. **Computational Animal Breeding**. Athens: University of Georgia. 1999. 190 p.

MISZTAL, I.; PEREZ-ENCISO, M. Sparse matrix inversion for restricted likelihood estimation of variance components by expectation-maximization. **Journal of Dairy Science**, v. 76, p. 1479-1483, 1993.

MONTAGNON, C.; FLORI, A.; CILAS, C. A new method to assess competition in coffee clonal trials with single-tree plots in Cote d'Ivoire. **Agronomy Journal**, v. 93, p. 227-231, 2001.

MOOD, A. M.; GRAYBILL, F. A. **Introduction to the theory of statistics**. New York: McGraw-Hill, 1963. 464 p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. Tokyo: McGraw-Hill, 1974. 564 p.

MOREAU, L.; MONOD, H.; CHARCOSSET, A.; GALLAIS, A. Marker assisted selection with spatial analysis of unreplicated field trials. **Theoretical and Applied Genetics**, v. 98, p. 234-242, 1999.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 2nd ed. Wallingford: CAB International, 2005. 368 p.

MRODE, R. A. Understanding cow evaluations in a random regression model. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte, Intituto Prociência, 2006. 1 CD-ROM.

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v. 124, p. 342-355, 2007.

MUIR, W.M. Incorporation of competitive effects in forest trees or animal breeding programs. **Genetics**, v. 170, p. 1247-1259, 2005.

MURILLO, O.; BADILLA, Y.; RESENDE, M. D. V. de. Quantitative genetics studies and selection of teak (*Tectona grandis*) in Costa Rica. **Tree Genetics and Genomes**, 2007. Submetido para publicação.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society, Series A**, v. 135, p. 370-384, 1972.

NETER, J.; KUTNER, M. H.; NACHTSHEIN, C. J.; WASSERMAN, W. **Applied linear statistical models**. Chicago: Irwin, 1996. 1408 p.

NEWMAN, D. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. **Biometrika**, v. 31, p. 20-30, 1939.

NOUY, B.; ASMADY, A.; LUBIS, R. Effets de competition a Nord-Sumatra dans le essais genetiques sur palmier a huile: consequences sur l'evaluation du materiel vegetal. **Oleagineux**, v. 45, p. 245-255, 1990.

NUNEZ-ANTON, V.; ZIMMERMANN, D. L. Modelling non-stationary longitudinal data. **Biometrics**, v. 56, p. 699-705, 2000.

OLIVEIRA, R. A.; RESENDE, M. D. V. de; DAROS, E.; BESPALHOK FILHO, J. C.; ZAMBON, J. L.; IDO, O. T.; WEBER, H.; KOEHLER, H. S. Genotypic evaluation and selection of sugarcane clones in three environments in the state of Paraná. **Crop Breeding and Applied Biotechnology**, v. 5, n. 4, p. 426-434, 2005.

PACHECO, C. A. P.; FERRÃO, M. A. G.; CRUZ, C. D.; VENCOVSKY, R. Proposal for applying combined selection to diallel analysis. **Brazilian Journal of Genetics**, v. 20, n. 2, p. 299-306, 1997.

PANSE, K.; SUKHATME, P. V. **Statistical methods for agricultural workers**. New Delhi: ICAR Publ., 1954. 349 p.

PAPADAKIS, J. Advances in the analysis of field experiments. **Communicationes d'Academie d'Athenes**, v. 59, p. 326-342, 1984.

PAPADAKIS, J. **Agricultural research**: principles, methodology, suggestions. Buenos Aires: Ed. Argentina, 1970.

PAPADAKIS, J. **Method statistique pour des experiences sur champ**. Thessalonique: Institut d'Amelioration des Plantes a Salonique. 1937. v. 23, p. 1-30.

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

PAWITAN, Y. **In all likelihood**: statistical modelling and inference using likelihood. Oxford: Clarendon Press. 2001. 528 p.

PEARCE, S. C. Experimenting with organisms as blocks. **Biometrika**, v. 44, p. 141-149, 1957.

PEARSON, K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. **Philosophical Transactions of the Royal Society of London, Section A**, v. 200, p. 1-66, 1903.

PERECIN, D.; BARBOSA, J. C. Uma avaliação de seis procedimentos para comparações múltiplas. **Revista de Matemática e Estatística**, v. 6, p. 95-103, 1988.

PEREIRA, B. B. Inferência verossimilhança. **Boletim da Associação Brasileira de Estatística**, v. 38, p. 31-42, 1997.

PIEPHO, H. P. Analysing genotype-environment interaction data by mixed models with multiplicative terms. **Biometrics**, v. 53, p. 761-767, 1997.

PIEPHO, H. P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. **Theoretical and Applied Genetics**, v. 89, p. 647-654, 1994.

PIEPHO, H. P. Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. **Theoretical and Applied Genetics**, v. 97, p. 195-201, 1998.

PIEPHO, H. P. Stability analysis using SAS. **Agronomy Journal**, v. 91, p. 154-160, 1999.

PIEPHO, H. P.; BUCHSE, A.; EMRICH, K. A hitchhiker's guide to mixed models for randomized experiments. **Journal of Agronomy & Crop Science**, v. 189, p. 310-322, 2003.

PIEPHO, H. P.; MOHRING, J. Best linear unbiased prediction of cultivar effects for subdivided target regions. **Crop Science**, v. 145, p. 1151-1159, 2005.

PIEPHO, H. P.; MOHRING, J. Selection in cultivar trials – is it ignorable? **Crop Science**, v. 146, p. 193-202, 2006.

PIEPHO, H. P.; MOHRING, J.; MELCHINGER, A. E.; BUCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, 2007 (In press).

PIEPHO, H. P.; WILLIAMS, E. R. A comparison of experimental designs for selection in breeding trials with nested treatment structure. **Theoretical and Applied Genetics**, v. 113, p. 1505-1513, 2006.

PIMENTEL-GOMES, F. **Curso de estatística experimental**. 12. ed. São Paulo: Nobel, 1987. 466 p.

PIMENTEL-GOMES, F. **Curso de estatística experimental**. Piracicaba: ESALQ, 1966. 466 p.

PIMENTEL-GOMES, F. O problema do tamanho das parcelas em experimentos com plantas arbóreas. **Pesquisa Agropecuária Brasileira**, v. 19, n. 12, p. 1507-1512, 1984.

PITHUNCHARURNLAP, M.; BASFORD, K. E.; FEDERER, W. T. Neighbour analysis with adjustment for interplot competition. **Australian Journal of Statistics**, v. 35, n. 3, p. 263-270, 1993.

PLETCHER, S. D.; GEYER, C. J. The genetic analysis of age-dependent traits: modelling a character process. **Genetics**, v. 153, p. 825-833, 1999.

POONI, H.S.; JINKS, J.L.; CORNISH, M.A. The causes and consequences of non-normality in predicting the properties of recombinant inbred lines. **Heredity***,* v. 38, p. 329-338, 1977.

QIAO, C. G.; BASFORD, K. E.; DELACY, I. H.; COOPER, M. Evaluation of experimental designs and spatial analysis in wheat breeding trials. **Theoretical and Applied Genetics**, v. 100, p. 9-16, 2000.

QUAAS, R. L.; POLLAK, E. J. Modified equations for sire models with groups. **Journal of Dairy Science**, v. 64, p. 1868-1872, 1981.

RAMALHO, M. A. P.; FERREIRA, D.F.; OLIVEIRA, A. C. de. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2000. 303 p.

RAMALHO, M. A. P.; SANTOS, J. B.; ZIMMERMANN, M. J. O**. Genética quantitativa no melhoramento de plantas autógamas**. Goiânia: Universidade Federal de Goiás, 1993. 271 p.

RAO, C.R. **Advanced statistical methods in biometric research**. New York: J. Wiley and Sons, 1952. 390 p.

RAO, P. S. **Variance components**: mixed models methodologies and applications. Boca Raton: CRC Press, 1997.350 p.

RESENDE, M. D. V. de. **Análise estatística de modelos mistos via REML/BLUP na experimentação em melhoramento de plantas perenes**. Colombo: Embrapa Florestas, 2000. 101 p. (Embrapa Florestas. Documentos, 47).

RESENDE, M. D. V. de. Avanços da genética biométrica florestal. In: BANDEL, G.; VELLO, N. A.; MIRANDA FILHO, J. B. (Ed.). **Encontro sobre temas de genética e melhoramento**: genética biometrica vegetal. **Anais**. Piracicaba: ESALQ, 1997. p. 20-46.

RESENDE, M. D. V. de. Correções nas expressões do progresso genético com seleção em função da amostragem finita dentro de famílias e populações e implicações no melhoramento florestal. **Boletim de Pesquisa Florestal**, v. 22/23, p. 61-77, jan./dez. 1991.

RESENDE, M. D. V. de. Delineamento de experimentos de seleção para a maximização da acurácia seletiva e progresso genético. **Revista Árvore**, v. 19, n. 4, p. 479-500, out./dez. 1995.

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

RESENDE, M. D. V. de. **Inferência bayesiana e simulação estocástica (amostragem de Gibbs) na estimação de componentes de variância e valores genéticos em plantas perenes**. Colombo: Embrapa Florestas, 2000b. 68 p. (Embrapa Florestas. Documentos, 46).

RESENDE, M. D. V. de. **Métodos estatísticos ótimos na análise de experimentos de campo**. Colombo: Embrapa Florestas. 2004. 65 p. (Embrapa Florestas. Documentos, 100).

RESENDE, M. D. V. de. **Seleção de genótipos de milho (*Zea mays* L.) em solos contrastantes***.* 230 f. 1989. Dissertação (Mestrado em Genética) – ESALQ, Piracicaba.

RESENDE, M. D. V. de. **Selegen–Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Colombo: Embrapa Florestas, 2007. 358 p.

RESENDE, M. D. V. de. **Software Selegen–REML/BLUP**. Colombo: Embrapa Florestas, 2002b. 67 p. (Embrapa Florestas. Documentos, 77).

RESENDE, M. D. V. de. **Predição de valores genéticos, componentes de variância, delineamentos de cruzamento e estrutura de populações no melhoramento florestal.** 1999. 434 f. Tese (Doutorado em Genética) - Setor de Ciências Biológicas, Universidade Federal do Paraná, Curitiba.

RESENDE, M. D. V. de; BARBOSA, M. H. P. **Melhoramento genético de plantas de propagação assexuada**. Colombo: Embrapa Florestas, 2005. 130 p.

RESENDE, M. D. V. de; BARBOSA, M. H. P. Selection via simulated individual BLUP based on family genotypic effects in sugarcane. **Pesquisa Agropecuária Brasileira**, v. 41, n. 3, p. 421-429, 2006.

RESENDE, M. D. V. de; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística,** São Paulo, v. 20, p. 39-65, 2002.

RESENDE, M. D. V. de; DIAS, L. A. S. Aplicação da metodologia de modelos mistos (REML/BLUP) na estimação de parâmetros genéticos e predição de valores genéticos em espécies frutíferas. **Revista Brasileira de Fruticultura**, v. 22, n. 1, p. 44-52, 2000.

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

RESENDE, M. D. V. de; DUDA, L. L. ; GUIMARÃES, P. R. B.; FERNANDES, J. S. C. Análise de modelos lineares mistos via inferência *Bayesiana*. **Revista de Matemática e Estatística**. São Paulo, v. 19, p. 41-70, 2001a.

RESENDE, M. D. V. de; FERNANDES, J. S. C. Análises alternativas envolvendo o procedimento BLUP e o delineamento experimental de blocos incompletos ou látice. **Revista de Matemática e Estatística**, v. 18, p. 103-124, 2000.

RESENDE, M. D. V. de; FERNANDES, J. S. C. Procedimento BLUP individual para delineamentos experimentais aplicados ao melhoramento florestal. **Revista de Matemática e Estatística**, v. 17, p. 89-107, 1999.

RESENDE, M. D. V. de; FERNANDES, J. S. C.; SIMEÃO, R. M. BLUP individual multivariado em presença de interação genótipo x ambiente para delineamentos experimentais repetidos em vários ambientes. **Revista de Matemática e Estatística**, v. 17. p. 209-228, 1999.

RESENDE, M. D. V. de; HIGA, A. R. Maximização da eficiência da seleção em testes de progênies de *Eucalyptus* através da utilização de todos os efeitos do modelo matemático. **Boletim de Pesquisa Florestal**, n. 28/29, p. 37-55, 1994.

RESENDE, M. D. V. de; HIGA, A. R.; LAVORANTI, O. J. Predição de valores genéticos no melhoramento de *Eucalyptus* – melhor predição linear. In: CONGRESSO FLORESTAL PANAMERICANO, 1.; CONGRESSO FLORESTAL BRASILEIRO, 7., 1993, Curitiba. **Floresta para o Desenvolvimento**: Política, Ambiente, Tecnologia e Mercado: anais. São Paulo: SBS; [S.l.]: SBEF, 1993. p. 144-147.

RESENDE, M. D. V. de; HIGA, A. R.; LAVORANTI, O. J. Regressão geno-fenotípica multivariada e maximização do progresso genético em programas de melhoramento de *Eucalyptus*. **Boletim de Pesquisa Florestal**, n. 28/29, p. 57-71, 1994.

RESENDE, M. D. V. de; OLIVEIRA, E. B. de. Sistema “SELEGEN” – Seleção Genética Computadorizada para o melhoramento de espécies perenes. **Pesquisa Agropecuária Brasileira**, v. 32, n. 9, p. 931-939, 1997.

RESENDE, M. D. V. de; OLIVEIRA, E. B. DE; MELINSKI, L. C.; GOULART, F. S.; OAIDA, G. R. **SELEGEN - Seleção Genética Computadorizada**: manual do usuário. Colombo: EMBRAPA-CNPF, 1994. 31 p.

RESENDE, M. D. V. de; PEREZ, J. R. H. R. **Genética e melhoramento de ovinos**. Curitiba: Ed. da Universidade Federal do Paraná, 2001. 186 p.

RESENDE, M. D. V. de; PEREZ, J. R. H. R. **Genética quantitativa e estatística no melhoramento animal**. Curitiba: Universidade Federal do Paraná, Imprensa Universitária, 1999. 494 p.

RESENDE, M. D. V. de; PRATES, D. F.; JESUS, A.; YAMADA, C. K. Estimação de componentes de variância e predição de valores genéticos pelo método da máxima verossimilhança restrita (REML) e melhor predição linear não viciada (BLUP) em *Pinus*. **Boletim de Pesquisa Florestal**, n. 32/33, p. 18-45, 1996.

RESENDE, M. D. V. de; RESENDE, R. M. S.; JANK, L.; VALLE, C. B. Experimentação e análise estatística no melhoramento de forrageiras. In: VALLE, C. B.; JANK, L.; RESENDE, R. M. S. **Melhoramento de forrageiras tropicais**. Campo Grande: Embrapa Gado de Corte. 2007. p. 1-97.

RESENDE, M. D. V. de; REZENDE, G. D. S. P.; FERNANDES, J. S. C. Regressão aleatória e funções de covariância na análise de medidas repetidas. **Revista de Matemática e Estatística,** São Paulo, v. 19, p. 21-40, 2001b.

RESENDE, M. D. V. de; STRINGER, J. K.; CULLIS, B. C; THOMPSON, R. Joint modelling of competition and spatial variability in forest field trials. **Brazilian Journal of Mathematics and Statistics**, v. 23, n. 2, p. 7-22, 2005.

RESENDE, M. D. V. de; STRINGER, J. K.; CULLIS, B. R.; THOMPSON, R. Joint modelling of competition and spatial variability in forest field trials. In: IUFRO Conference on Eucalyptus, 2004, Aveiro. **Proceedings**… Aveiro: Ed. Raíz, 2004. v. 1, p. 330-332.

RESENDE, M. D. V. de; STURION, J. A. Análise estatística espacial de experimentos via modelos mistos individuais com erros modelados por processos ARIMA em duas dimensões. **Revista de Matemática e Estatística**, v. 21, n. 1, p. 7-33, 2003.

RESENDE, M. D. V. de; STURION, J. A. **Analise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de series temporais empregando REML/BLUP ao nível individual**. Colombo: Embrapa Florestas, 2001. 80 p.

RESENDE, M. D. V. de; STURION, J. A.; MENDES, S. **Genética e melhoramento da erva-mate (*Ilex paraguariensis* St. Hill**). Colombo: EMBRAPA-CNPF, 1995. 33 p. (EMBRAPA-CNPF. Documentos, 25).

RESENDE, M. D. V. de; THOMPSON, R. Factor analytic multiplicative mixed models in the analysis of multiple experiments. **Brazilian Journal of Mathematics and Statistics**, v. 22, n. 2, p. 31- 52, 2004.

RESENDE, M. D. V. de; THOMPSON, R. **Multivariate spatial statistical analysis of multiple experiments and longitudinal data**. Colombo: Embrapa Florestas, 2003. 126 p. (Embrapa Florestas. Documentos, 90).

RESENDE, M. D. V. de; THOMPSON, R.; WELHAM, S. J.; BAIERL, A. Multivariate spatial statistical analysis in perennial crops. In: INTERNATIONAL BIOMETRIC SOCIETY CONFERENCE – BRITISH REGION, Reading. **Proceedings**… Reading: University of Reading, School of Applied Statistics, 2003. p. 70-71.

RESENDE, M. D. V. de; THOMPSON, R.; WELHAM, S. Multivariate spatial statistical analysis of longitudinal data in perennial crops. **Brazilian Journal of Mathematics and Statistics**, v. 24, n. 1, p. 147-169, 2006.

RESENDE, M. D. V. de; VENCOVSKY, R.; FERNANDES, J. S. C. Selection and genetic gains in populations of Eucalyptus with an mixed mating system. In: CRCTHF-IUFRO CONFERENCE. EUCALYPT PLANTATIONS: IMPROVING FIBRE YIELD AND QUALITY, 1995, Hobart. **Proceedings**... Hobart: CRCTHF, 1995. p. 191-193.

RIBEIRO JUNIOR, P. J. **Métodos geoestatísticos no estudo da variabilidade espacial de parâmetros do solo**. 1995. 99 f. Dissertação (Mestrado) – ESALQ, Piracicaba.

RIDGMAN, W. J. **Experimentation in biology**. Glasgow: Blackie, 1975. 233 p.

ROBBINS, H. The empirical Bayes approach to statistical decision problems. **Annals of Mathematical Statistics**, v. 35, n. 1, p. 1-20, 1964.

ROBERTSON, A. Experimental design in the evaluation of genetic parameters. **Biometrics**, Washington, v. 15, p. 219-226, 1959.

ROBERTSON, A. Weighting in estimation of variance components. **Biometrics**, v. 18, p. 413-415, 1962.

ROWE, S. J; WHITE, I. M. S.; AVENDANO, S.; HILL, W. G. Genetic heterogeneity of residual variance in broiler chickens. **Genetics Selection Evolution**, v. 38, p. 617-635, 2006.

RUGGIERO, M. A. G.; LOPES, V. R. L. **Cálculo numérico: aspectos teóricos e computacionais**. São Paulo: Makron Books, 1996.

SAMPAIO, I. B. M. **Estatística aplicada à experimentação animal**. Belo Horizonte: Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia, 1998. 221 p.

SANCRISTOBAL, M.; ROBERT-GRANIER, C.; FOULLEY, J. L. Hétéroscédasticité et modéles linéaires mixtes: théorie et applications en génétique quantitative. **Journal de la Société Française de Statistique**, v. 143, p. 155-165, 2002.

SCHAEFFER, L. R. Applications of random regression models in animal breeding. **Livestock Production Science**, v. 86, p. 35-45, 2004.

SCHAEFFER, L. R. **Linear models**. 1999. Disponível em: <www.**http://cgil.uoguelph.ca/people/faculty/lschaeffer.html>**. Acesso em: 15 jan. 2007.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 123, p. 218-223, 2006.

SCHAEFFER, L. R.; WILTON, J. W.; THOMPSON, R. Simultaneous estimation of variance and covariance components from multitrait mixed model equations. **Biometrics**, v.34, p.199-208, 1978.

SCHALL, R. Estimation in generalized linear models with random effects. **Biometrika**, v. 78, p. 719-727, 1991.

SCHEFFÉ, H. **The analysis of variance**. New York: J. Wiley, 1959. 477 p.

SCHNEEBERGER, M.; BARWICK, S. A.; CROW, G. H.; HAMMOND, K. Economic indices using breeding values predicted by BLUP. **Journal of Animal Breeding and Genetics**, v. 109, p. 180-187, 1992.

SCHOENBERG, I. J. Contributions to the problem of approximation of equidistant data by analitic functions. **Quarterly Applied Mathematics**, v. 4, p. 44-99, 1946.

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

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, v. 30, n. 3, p. 507-512, 1974.

SEARLE, S. R. Built in restrictions on best linear unbiased predictios (BLUP) of random effects in mixed models. **The American Statistician**, v. 51, n. 1, p. 19-21, 1997b.

SEARLE, S. R. C.R. Henderson, the statistician: his contributions to variance components estimation. **Journal of Dairy Science**, v. 74, p. 4035-4044, 1991.

SEARLE, S. R. **Linear models for unbalanced data**. New York: J. Wiley, 1987. 36 p.

SEARLE, S. R. **Linear models**. New York: J. Wiley, 1971.

SEARLE, S. R. Matrix algebra useful for statistics. New York. J. Wiley. 1982.

SEARLE, S. R. The matrix handling of BLUE and BLUP in the mixed linear model. **Linear Algebra and its Applications**, v. 264, p. 291-311, 1997a.

SEARLE, S. R.; CASELLA, G.; McCULLOCH, C. E. **Variance components**. New York: J. Wiley, 1992. 528 p.

SEVERINI, T. A. **Likelihood methods in statistics**. Oxford: Clarendon Press, 2001. 392 p.

SHAM, P. **Statistics in human genetics**. London: Arnold, 1998. 290 p.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, p. 591-611, 1965.

SHAW, R. G. Maximum-likelihood approaches to quantitative genetics of natural populations. **Evolution**, v. 41, p. 812-826, 1987.

SIEGEL, S. **Nonparametric statistics for the behavioral sciences**. Tokyo: McGraw-Hill, 1956. 312 p.

SILVA, E. C.; FERREIRA, D. F.; BEARZOTI, E. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. **Ciência e Agrotecnologia**, v. 23, n. 3, p. 687-696, 1999.

SILVEY, S. D. **Statistical inference**. 2. ed. London: Chapman & Hall, 1975. 191 p.

SMITH, A. F. M. Bayesian statistics. Present position and potencial developments: some personal views. **Journal of the Royal Statistical Society**, **Series A**, v. 147, p. 245-259, 1984.

SMITH, A.; CULLIS, B. R.; THOMPSON, R. Analysing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. **Biometrics**, v. 57, p. 1138-1147, 2001.

SMITH, A.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. **Journal of Agricultural Sciences**, v. 143, p. 1-14, 2005.

SMITH, A.; CULLIS, B.; GILMOUR, A. The analysis of crop variety evaluation data in Australia. **Australian and New Zealand Journal of Statistics**, v. 43, n. 2, p. 129-145, 2001.

SMITH, S. P.; GRASER, H. U. Estimating variance components in a class of mixed models by restricted maximum likelihood. **Journal of Dairy Science**, Champaign, v. 69, p. 1156-1165, 1986.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**. 6. ed. Iowa: Iowa State University Press, 1967. 507 p.

SOLBERG, T. R.; SONESSON, A.; WOOLIAMS, J.; MEUWISSEN, T. H. E. Genomic selection using different marker types and density. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

SORENSEN, D; GIANOLA, D. **Likelihood, bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag, 2002. 740 p.

SOUZA, J. **Análise de componentes principais; métodos estatísticos nas ciências psicossociais**. Brasília, DF: Thesaurus. 1998. 67 p.

STEEL, R. G. D.; TORRIE, J. H. **Principles and procedures of statistics**. 1th. ed. New York: Mac Graw-Hill, 1960. 633 p.

STEIN, C. A two sample test for a linear hypothesis whose power is independent of the variance. **Annals of Mathematical Statistics**, v. 16, p. 243-258, 1945.

STEIN, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 3., 1955, Berkeley. **Proceedings**... Berkeley: University of California Press, 1955. p. 197-206, 1955.

STIGLER, S. M. Who discovered the Bayesian theorem?. **American Statistician**, v. 37, p. 290-296, 1983.

STORCK, L; GARCIA, D. C.; LOPES, S. J.; ESTEFANEL, V. **Experimentação vegetal**. Santa Maria: Ed. da Universidade Federal de Santa Maria, 2000. 199 p.

STRAM, D. O.; LEE, J. W. Variance components testing in longitudinal mixed effects setting. **Biometrics**, v. 50, p. 1171-1177, 1994.

STRINGER, J. K.; CULLIS, B. R. Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. **Australian Journal of Agricultural Research**, v. 53, n. 8, p. 911-918, 2002a.

STRINGER, J. K.; CULLIS, B. R. Joint modelling of spatial variability interplot competition. In: AUSTRALASIAN PLANT BREEDING CONFERENCE, 12, Perth. **Proceedings**, Perth: Australian Plant Breeding Asscіation, 2002b. p. 614-619.

STROUP, W. W.; MULITZE, D. K. Nearest neighbour adjusted best linear unbiased prediction. **American Statistician**, v. 45, p. 194-200, 1991.

TAKAHASHI, K.; FAGAN, J.; CHIN, M. S. Formation of a sparse bus impedance matrix and its application to short circuit study. In: Institutional Pica Conference, 8, 1973. **Proceedings** Minneapolis: IEEE Power Engineering Society, 1973. p.63.

TALBOT, M.; MILNER, A. D.; NUTKINS, M. A. E.; LAW, J. R. Effect of interference between plots on yield performance in crop variety trials. **Journal of Agricultural Science**, Cambridge, v. 124, p. 335-342, 1995.

THEIL, H. **Principles of econometrics**. New York: John Wiley & Sons, 1971.

THOMPSON JUNIOR, W. A. The problem of negative estimates of variance components. **Annals of Mathematical Statistics**, v. 33, p. 273-284, 1962.

THOMPSON, R. A note on restricted maximum likelihood estimation with an alternative outlier model. **Journal of the Royal Statistics Society**, **Series B**, v. 47, p. 53-55, 1985.

THOMPSON, R. A review of genetic parameter estimation. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**. Paris: INRA, 2002. p. 19-23.

THOMPSON, R. Analysis of cereal variety trials. EU HARMA WORKSHOP, 7., 1996, Dublin. **Proceedings**... Dublin: [s.n.], 1996. p. 7-10. Editado por: Connolly, J.; Williams, D.

THOMPSON, R. Estimation of genetic parameters. In: MRODE, R. A. (Ed.) **Linear models for the prediction of animal breeding values**. 2. ed. Wallingford: CAB International, 2005. p. 235-245.

THOMPSON, R. Estimation of quantitative genetic parameters. In:INTERNATIONAL CONFERENCE ON QUANTITATIVE GENETICS, 1977, Ames. **Proceedings**. Ames: Iowa State University Press. p. 639-657. Editor: O. Kempthorne.

THOMPSON, R. Iterative estimation of variance components for non-orthogonal data. **Biometrics**, v. 25, p. 767-773, 1969.

THOMPSON, R. Maximum likelihood estimation of variance components. **Mathematik Operationsforsh Statistik**, v. 11, p. 545-561, 1980.

THOMPSON, R. Relationship between the cumulative difference and best linear unbiased predictor methods of evaluating bulls. **Animal Production**, v. 23, p. 15-24, 1976.

THOMPSON, R. Sire evaluation. **Biometrics**, v. 35, p. 339-353, 1979.

THOMPSON, R. The estimation of heritability with unbalanced data. **Biometrics**, v. 33, p. 485-504, 1977.

THOMPSON, R. The estimation of variance and covariance components when records are subject to culling. **Biometrics**, v. 29, p. 527-550, 1973.

THOMPSON, R.; CULLIS, B. R.; SMITH, A. B.; GILMOUR, A. R. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. **Australian and New Zealand Journal of Statistics**, v. 45, n. 4, p. 445-459, 2003.

THOMPSON, R.; MEYER, K. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. **Livestock Production Science**, v. 15, p. 299-313, 1986.

THOMPSON, R.; WELHAM, S. J. REML analysis of mixed models. In: PAYNE, R. (Ed.). **GenStat 6 Release 6.1.**: the guide to GenStat, v. 2 – Statistics. Harpenden: Rothamsted Research, 2003. p. 469-560.

THOMPSON, R; BROTHERSTONE, S.; WHITE, M. S. Estimation of quantitative genetic parameters. **Philosophical Transaction of the Royal Society of Britain**, v. 360, p. 1469-1477, 2005.

THOMPSON, R; WRAY, N. R.; CRUMP, R. E. Calculation of prediction error variances using sparse matrix methods. **Journal of Animal Breeding and Genetics**, v. 111, p. 102-109, 1994.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistics Society Series B**, v. 58, p.267-288, 1996.

TUKEY, J. W. **Exploratory data analysis**. New York: Addison-Wesley, 1977.

TUKEY, J. W. One degree of freedom for non-additivity. **Biometrics**, v. 5, p. 232-242, 1949.

VALEN, L. van. The statistics of variation. **Evolution Theory**, v. 4, p. 33-43, 1978.

VALENTE, J. M. G. P. **Geomatemática**: lições de geoestatística. 2. ed. Ouro Preto: Fundação Gorceix, 1989. v. 3.

VENCOVSKY, R. Effective size of monoecious populations submitted to artificial selection. **Brazilian Journal of Genetics**, v. 1, n. 3, p. 181-191, 1978.

VENCOVSKY, R. Genética quantitativa. In: KERR, W. E. (Ed.). **Melhoramento e genética**. São Paulo: Melhoramentos, 1969. p. 17-38

VENCOVSKY, R. Herança quantitativa. In: PATERNIANI, E.; VIEGAS, G. P. (Ed.) **Melhoramento e produção de milho**. 2. ed. Campinas: Fundação Cargill, 1987a. v. 1, p. 137-214.

VENCOVSKY, R. Repetibilidade. In: VENCOVSKY, R. **Princípios de genética quantitativa**. Piracicaba: ESALQ, 1972. p. 47-52,

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. Ribeirão Preto: Sociedade Brasileira de Genética, 1992. 486 p.

VENCOVSKY, R.; PEREIRA, M. B.; CRISOSTÓMO, J. R.; FERREIRA, M. A. J. F. Genética e melhoramento de populações mistas.. In: NASS, L. L.; VALOAIS, A. C. C.; MELO, I. S.; VALADARES-INGLIS, M. C. (Org.). **Recursos genéticos e melhoramento**. Rondonópolis: Fundação MT, 2001. p. 231-281.

VERBYLA, A. P; CULLIS, B. R.; KENWARD, M. G.; WELHAM, S. J. The analyses of designed experiments and longitudinal data using smootthing splines. **Journal of the Royal Statistics Society**, **Series C,** v. 48, p. 269-311, 1999.

VERNEQUE, R. S.; VALENTE, J. Avaliação genética de vacas e touros. In: VALENTE, J.; DURÃES, M.C.; MARTINEZ, M.L.; TEIXEIRA, N.M. (Ed.). **Melhoramento genético de bovinos de leite**. Juiz de Fora: Embrapa Gado de Leite, 2001. v. 1, p. 127-154.

VESSEREAU, A. **Méthodes statistiques en biologie et en agronomie**. Paris: Baillière, 1960.

VIANELLI, S. **Metodologia statistica delle scienze agrarie**. Bologna: Edizioni Agricole. 1954.

VIEIRA, S. **Estatística experimental**. São Paulo: Atlas, 1999, 185 p.

VINOD, H. D. Simulation and extension of a minimum mean squared error estimator in comparison with Stein´s. **Technometrics**, v.18, n. 4, p. 491-496, 1976.

VISSCHER, P. M.; GODDARD, M. E. Fixed and random contemporary groups. **Journal of Dairy Science**, v. 76, n. 5, p. 1444-1454, 1993.

VISSCHER, P. M.; HILL, W. G. Heterogeneity of variance and dairy cattle breeding**. Animal Production**, v. 55, p. 321-329, 1992.

VLECK, L. D. van. **Selection index and introduction to mixed model methods**. Boca Raton: CRC Press, 1993a. 512 p.

VLECK, L. D; CASSADY, J. P. Unexpected estimates of variance components with a true model containing genetic competition effects. **Journal of Animal Science**, v. 83, p. 68-74, 2005.

VLECK, L.D. van; POLLAK, E. J.; OLTENACU, E. A. B. **Genetics for the animal sciences**. New York : W.H. Freeman, 1987. 391 p.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical Society**, v. 54, p. 426-482, 1943.

WALSH, J. L.; AHLBERS, J. H.; NILSON, E. N. Best approximation properties of the spline fit. **Journal of Mathematical Mechanics**, v. 11, p. 225-234, 1962.

WEEKS, D. L.; WILLIAMS, D. R. A note on the determination of connectedeness in an N-way cross classification. **Technometrics**, v. 6, p. 319-324, 1964.

WEIGEL, K. A., GIANOLA, D.; TEMPELMAN, R. J.; MATOS, C. A.; CHEN, I. H. C. Improving estimates of fixed effects in a mixed linear model. **Journal of Dairy Science**, v. 74, p. 3174-3182, 1991.

WELHAM, S. J.; CULLIS, B. R.; GOGEL, B. J.; GILMOUR, A. R.; THOMPSON, R. Prediction in linear mixed models. **Australian and New Zealand Journal of Statistics,** v. 46, p. 325-347, 2004.

WELHAM, S. J.; CULLIS, B. R.; KENWARD, M.; THOMPSON, R. A comparison of mixed model splines for curve fitting. **Australian and New Zealand Journal of Statistics**, v. 49, p. 1-23, 2007.

WELHAM, S. J.; CULLIS, B. R.; KENWARD, M.; THOMPSON, R. The analysis of longitudinal data using mixed model L-splines", **Biometrics**, v. 62, p. 392-401, 2006.

WELHAM, S. J.; THOMPSON, R. Likelihood ratio tests for fixed models terms using residual maximum likelihood. **Journal of the Royal Statistical Society, Series B**, v. 59, p. 701-719, 1997.

WELHAM, S. J.; THOMPSON, R.; GILMOUR, A. R. A general form for specification of correlated error models with allowance for heterogeneity. In: SYMPOSIUM ON COMPUTATIONAL STATISTICS, 3., 1998, Bristol. **Proceedings in computational statistics**. Heidelberg: Physica-Verlag, 1998. p. 479-484. COMPSTAT'98.

WHITE, I. M. S.; ROEHE, R.; KNAP, P. W.; BROTHERSTONE, S. Variance components for survival of piglets at farrowing using a reduced animal model. **Genetics, Selection, Evolution**, v. 38, p. 359-370, 2006.

WHITE, I. M. S.; THOMPSON, R.; BROTHERSTONE, S. Genetic and environmental smoothing of lactation curves with cubic splines. **Journal of Dairy Science**, v. 82, p. 632-638, 1999.

WILKINSON, G. N.; ECKERT, S.R .; HANCOCK, T.W.; MAYO, O. Nearest neighbor (NN) analysis of field experiments. **Journal of the Royal Statistical Society, Series B**, v. 45, p. 151-211, 1983.

WILKS, S. S. The large sample distribution of the likelihood ratio for testing composite hypothesis. **Annals of Mathematical Statistics**, v. 9, p. 60-62, 1938.

WILLIAMS, E. R. A neighbor model for field experiments. **Biometrika**, v. 73, p. 279-287, 1986.

WILLIAMS, J. S. The evaluation of a selection index. **Biometrics**, v. 18, p. 375-393, 1962.

WRIGHT, J. W.; PAULEY, S. S.; POLK, R. B; JOKELA, J. J. Performance of Scotch pine varieties in North Central Region. **Silvae Genetica**, v. 15, p. 101-110, 1966.

WU, R; MA, C. X.; PAINTER, I.; ZENG, Z. B. Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. **Theoretical Population Biology**, v. 61, p. 349-363, 2002.

YATES, F. A new method of arranging variety trials involving a large number of varieties. **Journal of Agricultural Sciences**, v. 26, p. 424-455, 1936.

YATES, F. The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association**, v. 29, p. 51-66, 1934.

YATES, F. The recovery of inter-block information in balanced incomplete block designs**. Annals of Eugenics**, v. 10, p. 317-325, 1940.

ZEGER, S. L.; LIANG, K. Y. Longitudinal data analysis for discrete and continuous outcomes. **Biometrics**, v. 42, p. 121-130, 1986.

ZEGER, S. L.; LIANG, K. Y.; ALBERT, P. S. Models for longitudinal data: a generalized estimation approach. **Biometrics**, v. 44, p. 1049-1060, 1988.

ZIMMERMAN, D.I.; HARVILLE, D.A. A random field approach to the analysis of field-plot experiments and other spatial experiments. **Biometrics**, v. 47, p. 223-239, 1991.

ZOLLENKOPF, K. Bi-Factorisation - Basic computational algorithm and programming techniques. In: REID, J. K. (Ed.). **Large sparse sets of linear equations**. London: Academic Press, 1971. p. 75-96.