Métodos Estatísticos Ótimos na Análise de Experimentos de Campo no Melhoramento de Plantas

Marcos Deon Vilela de Resende

1. Análise Estatística de Experimentos de Campo

A experimentação de campo apresenta certas peculiaridades, dentre as quais destacam-se: (i) 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; (ii) tendência ambiental ou variabilidade ambiental à pequena escala devido a fatores do solo tais como fertilidade, umidade, dentre outros; (iii) competição entre genótipos devido a diferentes agressividades e sensibilidades dos diferentes materiais genéticos; (iv) heterogeneidade de variâncias entre experimentos de uma rede experimental ou entre medidas repetidas tomadas em uma mesma unidade experimental.

Para contornar o problema exposto em (i) deve-se adotar o procedimento REML/BLUP (máxima verossimilhança residual ou restrita/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 estes efeitos forem considerados fixos. Estes 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. 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 tratamento. Em geral, os modelos de competição/interferência são aplicados em conjunto com a análise espacial via o procedimento REML. A questão (iv) relativa à heterogeneidade de variâncias pode ser contornada pelo próprio REML, o qual permite considerar variâncias heterogêneas para os vários efeitos do modelo sob análise.

Este trabalho tem como objetivo apontar alguns procedimentos ótimos de análise estatística de dados advindos de experimentos de campo. Inúmeros procedimentos estatísticos existem para esta finalidade mas muitos deles não são ótimos para uso de forma generalizada. Muitas vezes o usuário aplica, ao mesmo conjunto de dados experimentais, um grande número de métodos disponíveis sem se atentar para a escolha do procedimento que é ótimo por concepção. Assim, são apontados aqui procedimentos ótimos especialmente relacionados ao melhoramento genético, enfatizando-se os métodos BLUP e IME para a predição de valores genéticos; REML para a estimação de componentes de variância; modelos fator analíticos multiplicativos mistos (FAMM) para a análise de múltiplos experimentos e da divergência genética, adaptabilidade e estabilidade genotípica; média harmônica da performance relativa dos valores genéticos (MHPRVG) para inferência simultânea sobre produtividade, estabilidade e adaptabilidade; modelos autoregressivos separáveis de primeira ordem em duas dimensões (linha e coluna) para análise espacial de experimentos (AR1 x AR1); modelos de competição genotípica para contemplar a interferência entre tratamentos; modelo auto-regressivo com variâncias heterogêneas (ARH) ou modelo ante-dependência estruturado (SAD) para análise de medidas repetidas com correlações de magnitudes decrescentes com o aumento da distância entre medições;

modelos para estudos de QTL envolvendo simultaneamente os efeitos dos QTL marcados, dos QTL não marcados, da dependência espacial e da competição entre genótipos. Grande parte do material apresentado aqui baseia-se no trabalho de Resende (2004)

2. Uso do REML em detrimento da ANOVA

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 como suposição básica a independência dos erros. O método REML permite relaxar esta suposição de independência permitindo maior flexibilidade na modelagem. Tal procedimento foi criado pelos pesquisadores ingleses Desmond Patterson e Robin Thompson 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 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 problema central do melhoramento genético é a predição dos valores genéticos dos vários candidatos à seleção. E esta predição necessita de componentes de variância conhecidos ou estimados com precisão. O procedimento ótimo de predição de valores genéticos é o BLUP e o procedimento ótimo de estimação de componentes de variância é o REML. Estes procedimentos estão associados a um modelo linear misto, isto é, modelo que contém efeitos fixos além da média geral e efeitos aleatórios além do erro. Nos experimentos de campo, os efeitos de tratamentos e os efeitos ambientais identificáveis (blocos, locais) podem ser considerados fixos ou aleatórios. Neste caso, para se ter um modelo misto, um destes efeitos deve ser considerado fixo e o outro aleatório. Assim, são modelos mistos: (a) modelos com efeitos aleatórios de tratamentos e efeitos fixos de ambiente; (b) modelos com efeitos fixos de tratamentos e efeitos aleatórios de ambiente. Inferências sobre os tratamentos são realizadas pelos procedimentos REML/BLUP para a situação (a) e REML/GLS para a situação (b). Em ambos os tipos de modelo, as inferências sobre tratamentos podem ser extrapoladas para toda a população de efeitos ambientais (locais, blocos), pois a interação tratamentos x ambientes será de efeito aleatório para ambos os modelos, o que é essencial.

Um fator é comumente tomado como aleatório se os níveis observados podem ser assumidos como uma amostra aleatória de uma população, ou seja, amostra aleatória de uma assumida distribuição de probabilidade (distribuição normal para aplicação do REML). Em melhoramento genético, a consideração dos efeitos de tratamentos como aleatório conduz a maior acurácia preditiva. Isto se deve ao fato de que as predições dos efeitos aleatórios são forçadas (shrinkage) em direção à média geral, penalizando predições baseadas em pequenas amostras. Isto não ocorre quando os efeitos são considerados fixos. Os efeitos ambientais podem ser considerados fixos (quando completos em termos da representatividade dos efeitos de tratamentos em cada um de seus níveis) ou aleatórios (quando incompletos). Tanto no caso dos efeitos de tratamentos quanto de ambiente, a suposição de fatores aleatórios depende também de um razoável número de níveis em cada fator (usualmente maior que 10) e de uma razoável simetria na distribuição (normalidade). Com a aceitação dos efeitos de tratamentos como aleatório, os testes de comparações múltiplas de médias de tratamentos são reprováveis. Em todos os ensaios de comparação de materiais genéticos com 10 ou mais acessos ou entradas deve-se usar preferencialmente o

REML/BLUP (tratamentos como efeitos aleatórios) e não testes de comparação de médias (tratamentos como efeitos fixos).

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 números de tratamentos menores que 10. A consideração dos efeitos de tratamentos como aleatórios é essencial ao melhoramento genético. É a única forma de se fazer seleção genética. Caso contrário, a seleção é fenotípica e não genética. Isto porque a única forma de se eliminar os efeitos ambientais residuais embutidos nos dados fenotípicos é por meio do shrinkage ou multiplicação do valor fenotípico corrigido por uma função da herdabilidade do caráter sob seleção. Outros autores (Hill & Rosenberger, 1985; Stroup & Mulitze, 1991; Piepho, 1994; Resende et al. 1996; Piepho, 1998; Smith, Cullis & Gilmour, 2001; Duarte, 2000; Resende, 2002a) também enfatizam a necessidade de se considerar os materiais genéticos como de efeitos aleatórios, mesmo que os materiais sejam considerados de efeitos fixos em outras abordagens padrão. Os efeitos ambientais (blocos, locais) podem ser considerados fixos ou aleatórios dependendo da situação, mas os efeitos genéticos devem ser considerados aleatórios. Assim, os modelos em melhoramento genético devem sempre ser aleatórios ou misto com genótipos aleatórios.

As vantagens do uso de modelos mistos (REML) em relação ao uso de modelos completamente fixos (ANOVA) são: (a) produzem estimativas ou predições mais acuradas de efeitos de tratamentos quando existem dados perdidos nos experimentos; (b) as predições dos efeitos aleatórios são forçadas (shrinkage) em direção à média geral, penalizando estimativas baseadas em pequenas amostras; (c) permitem o ajuste de diferentes variâncias para cada grupo de tratamentos, ou seja, permite considerar variâncias heterogêneas; (d) resolvem o problema de estimação quando se tem dados perdidos; (e) a modelagem da estrutura de correlação em experimentos com dependência espacial, medidas repetidas e em múltiplos experimentos conduz a estimativas mais precisas (o modelo de efeitos fixos assume que todas as observações são não correlacionadas); (f) os resultados são mais apropriados para a inferência requerida quando a estrutura dos dados é hierárquica ou em multi-níveis. Como única desvantagem relata-se o maior número de suposições distribucionais que são feitas. É importante dizer que o procedimento REML pode ser derivado também sob o enfoque Bayesiano, fato que caracteriza a sua generalidade como procedimento ótimo.

Especificamente no caso dos modelos mistos com efeitos aleatórios de tratamentos, as propriedades do BLUP para os tratamentos são: maximização da acurácia seletiva, minimização do erro de predição, predição não viciada de valores genéticos, maximização do ganho genético por ciclo de seleção, maximização da probabilidade de selecionar o melhor entre dois genótipos; maximização da probabilidade de selecionar o melhor entre vários genótipos. É importante relatar também que o BLUP é o mais eficiente índice de seleção em termos de uso das informações de parentes, pois usa todas as informações disponíveis (efeitos aleatórios do modelo estatístico) assim como o faz o índice multi-efeitos (IME) relatado por Resende & Higa (1994).

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. No caso de dados desbalanceados, a ANOVA conduz a imprecisas estimativas de componentes de variância e consequentemente

a inacuradas predições de valores genéticos. Um software de fácil aplicação prática, destinado a aplicação corriqueira no melhoramento genético é o Selegen-REML/BLUP (Resende, 2002b). Um programa mais complexo e completo destinado a modelagem linear e linear generalizada via REML é o ASREML (Gilmour, Thompson, Cullis and Welham, 2002).

Para o caso de dados balanceados, o procedimento do índice multi-efeitos (Resende & Higa, 1994) é BLUP para a seleção de indivíduos em plantas perenes e o índice $\hat{g} = h_g^2(\bar{Y}_{i..} - \bar{Y}_{...})$ é BLUP para a seleção de materiais genéticos (híbridos, linhagens, progênies, clones, variedades, cultivares, populações) em plantas anuais e perenes, em que $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2/b + \sigma_\delta^2/nb)$, $\bar{Y}_{...}$ é a média geral do experimento, $\bar{Y}_{i..}$ é a média do material genético i avaliado em b repetições com n plantas por parcela. Os componentes de variância referem-se a variação entre materiais genéticos (*g*), entre parcelas (*c*) e dentro de parcelas (δ). É importante verificar que o fator de regressão ou shrinkage é o h_g^2. Em casos de dados desbalanceados existe um h_{gi}^2 para cada tratamento genético i o qual é função do tamanho amostral (n e b) associado a cada material genético. Além deste shrinkage diferenciado, o BLUP considera o desbalanceamento também por ocasião da correção para os efeitos fixos.

É importante relatar que um modelo completamente aleatório (com exceção da média geral) pode ser também submetido à análise REML/BLUP, obtendo-se todas as vantagens mencionadas. Pode-se dizer, adicionalmente que, nesse caso, o procedimento REML se assemelha ao procedimento de máxima verossimilhança não restrita (ML), aplicável a dados desbalanceados.

3. 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 (Resende, 2002a).

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 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 & 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:

$$I = b_1\delta_{ijk} + b_2 g_i + b_3\ c_{ij}$$
$$= b_1\left(Y_{ijk} - \overline{Y}_{ij\cdot}\right) + b_2\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}\ldots\right) + b_3\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}\ldots\right)$$

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:

$$b_1 = \frac{(1/2)\ \sigma_a^2}{\sigma_\delta^2}$$

$$b_2 = \frac{(2nb+1)}{2nb}\sigma_a^2$$
$$\overline{\sigma_p^2 + \sigma_c^2/b + \sigma_\delta^2/nb}$$

$$b_3 = \frac{[(1/2)/n]\ \sigma_a^2}{\sigma_c^2 + \sigma_\delta^2/n}$$

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 (r) segregantes. Neste caso, o BLUP para seleção pelos efeitos genéticos aditivos está associado ao seguinte índice multi-efeitos:

$$I_2 = b_1\delta_{ijkl} + b_2 g_i + b_3\ c_{ijl} + b_4 r_l$$
$$= b_1\left(Y_{ijkl} - \overline{Y}_{ij\cdot\cdot}\right) + b_2\left(\overline{Y}_{i\cdot\cdot\cdot} - \overline{Y}\ldots.\right) + b_3\left(\overline{Y}_{ij\cdot\cdot} - \overline{Y}_{i\cdot\cdot\cdot} - \overline{Y}_{\cdot j\cdot\cdot} + \overline{Y}\ldots.\right) + b_4\left(\overline{Y}_{\cdot\cdot l} - \overline{Y}\ldots.\right)$$, em que

$$b_4 = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_p^2/p + \sigma_c^2/bp + \sigma_\delta^2/nbp}.$$

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 $b_6 = \frac{[2n+1)/(2n)]\ \sigma_a^2}{\sigma_p^2 + \sigma_\delta^2/n}$. É importante relatar que neste caso a seleção não é puramente genética, pois o experimento não teve repetição, fato que prejudica também a casualização.

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 $\hat{g}_j > \hat{g}_k(1 - h^2_{gd})/(1 - h^2_{gd}) + (h^2_{gd} y - h^2_{gd} y + \hat{u} - \hat{u})/(h^2_{gd})$, 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 & Barbosa, 2004). 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 sequencial 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 no máximo 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, o n_j pode ser tomado como no máximo 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)*.

4. Uso do REML/GLS na 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 (com 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 (Litell, 2002). Após a estimação via REML, os testes de hipóteses e comparação de tratamentos podem ser realizados a semelhança do que é realizado tradicionalmente em associação com a ANOVA e com a estatística experimental tradicional.

5. Análise Estatística da Variabilidade Espacial em Fertilidade

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 existe e contribui para o aumento da variação residual, de forma que é 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 geoestatísticos e aos métodos de análise de séries temporais. 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 & 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. Assim, o vetor de erros é particionado em $e = \xi + \eta$. Os modelos de análise tradicionais não incluem o componente ξ. A variância dos resíduos é dada por $\mathrm{Var}(e) = \mathrm{Var}(\xi + \eta) = R = \Sigma = \sigma_\xi^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)] + I\sigma_\eta^2$, 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 $H = [\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$.

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.

Devido ao fato de que 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 & Verbyla, 1997) formalmente a equivalência entre a modelagem geoestatística exponencial e o modelo separável AR1 x AR1 para experimentos de campo. Também, o modelo linear de campo aleatório de Zimmerman & Harville é equivalente ao ajuste do modelo AR1 x AR1 sem o erro independente η. 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. Os modelos ARIMA englobam também as demais metodologias de análise de vizinhança apresentadas na literatura, tais quais os métodos de Papadakis e modificações (Gilmour, Thompson & Cullis, 1995).

Considerando um modelo com erros espacialmente correlacionados, tem-se o modelo $Y_{ijk} = \mu + p_i + b_j + c_{ij} + \xi_{ijk} + \eta_{ijk}$, em que *p*, *b*, *c*,$\xi$ e η são os efeitos de progênie, bloco (fixo), parcela e erros correlacionado e independente dentro de parcela, respectivamente. O índice multi-efeitos espacial (IMEE) para a predição dos efeitos genéticos aditivos é dado por $\hat{a}_{ijK} = b_1 p_i + b_2 c_{ij} + b_3 \xi_{ijk} + b_4 \eta_{ijk}$. 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 & Higa (1994). O modelo acima, em versão matricial é dado por

$$\begin{aligned} y &= X\beta + Za + Wc + \varepsilon \\ &= X\beta + Za + Wc + \xi + \eta \end{aligned}$$

, em que *a* e *c* são os efeitos aleatórios genéticos e de parcelas, respectivamente.

O IMEE em sua versão BLUP é dado por

$$\begin{bmatrix} X'X & X'Z & X'W & X'I \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'W & Z'I \\ W'X & W'Z^* & W'W + I\lambda_2 & W'I \\ I'X & I'Z^* & I'W & I'I + H^{-1}\lambda_3 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{a} \\ \tilde{c} \\ \tilde{\xi} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ I'y \end{bmatrix}$$

, 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} advem 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 herdablidade 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 $\sigma_\eta^2/\sigma_\varepsilon^2$ referentes aos modelos espacial e não espacial; (ii) 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 ($\lambda_1 = (1 - \hat{h}_{adj}^2)/\hat{h}_{adj}^2$), 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 & 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 $a = Z\tau + NZ\phi$ (ver tópico 4). A análise espacial pode ser útil também no aumento da eficiência da seleção massal. 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 correlacionado, após correção para os efeitos de linhas, colunas e $\hat{\xi}$.

Um outro procedimento melhorado em relação ao índice multi-efeitos tradicional de Resende & 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. O IMELC para a predição dos efeitos genéticos aditivos é dado por $\hat{a}_{(i)jKl} = b_1 p_{(i)} + b_2 l_j + b_3\kappa_k + b_4 c_{jk} + b_5\delta_{(i)jkl}$.

$$\hat{a}_{(i)jKl} = b_1\left(\bar{Y}_{(i)} - \bar{Y}_{...}\right) + b_2\left(\bar{Y}_{j..} - \bar{Y}_{...}\right) + b_3\left(\bar{Y}_{.k.} - \bar{Y}_{...}\right) + b_4\left(\bar{Y}_{jk.} - \bar{Y}_{(i)} - \bar{Y}_{j..} + \bar{Y}_{.k.} + 2\bar{Y}_{...}\right) + b_5\left(Y_{(i)jkl} - \bar{Y}_{jk.}\right)$$

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 (eficiência 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 & Welham, 2003). No caso, cada bloco continha 6 linhas (devido a 6 plantas por parcela). Com o DLC, o tamanho de cada bloco passou a equivaler a 1 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.

6. Análise Estatística da Interferência entre Tratamentos e Competição

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 diretamente 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, denominado interferência. Ajustamento para estes dois efeitos tendem a reduzir vícios e a melhorar a análise de experimentos de campo.

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 sensíveis, as quais tem 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. 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 e efeito indireto nos vizinhos (Stringer & Cullis, 2002; Resende, Stringer, Cullis & Thompson, 2004).

Estudo comparativo entre estes modelos via REML/BLUP em associação com a análise espacial revelaram: (i) o modelo fenotípico de competição, ajustado via covariável, é inconsistente e inadequado, visto que a covariável participa ao mesmo tempo da média e variância de y; (ii) o modelo fenotípico ajustado via perfil de verossimilhança é adequado pois contorna a dependência mencionada; (iii) o modelo de competição fenotípica, ajustado via perfil de verossimilhança, engloba todo o padrão de correlação, incluindo o efeito genético de competição e um balanço entre o efeito residual de competição e tendência ambiental espacial. (iv) com o uso do modelo fenotípico de competição ajustado via perfil

de verossimilhança, não há necessidade de inclusão dos efeitos de tendência espacial (análise com erros espaciais); (v) a análise espacial e o modelo fenotípico de competição diferem apenas em presença de competição em nível genético, portanto, sem competição genética o modelo de análise espacial (auto-regressivo) é suficiente; (vi) um modelo incluindo simultaneamente a interferência via o modelo genotípico de competição e a tendência espacial via um modelo auto-regressivo mostrou-se ideal em qualquer situação; (vii) o modelo simultâneo para competição genética e tendência em fertilidade considera explicitamente a competição em nível genético e também permite a covariância entre os efeitos genéticos diretos e indiretos sobre os vizinhos, sendo por isto preferível em relação ao modelo de competição fenotípica, o qual considera apenas implicitamente a competição genética (Resende, Stringer, Cullis & Thompson, 2004).

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 e é 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 dendezeiro.

Modelos com competição geralmente conduzem a menores estimativas de herdabilidade e ganhos genéticos estimados, mas são mais realísticos. O modelo apresentado é o mais completo que pode existir para um experimento de campo. Em algumas situações, alguns efeitos não são significativos e podem ser retirados do modelo, que converte-se então em modelos mais tradicionais (Resende, Stringer, Cullis & Thompson, 2004).

7. Análise Estatística 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. Os procedimentos de análise 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 & Wilkinson, o de Eberhart & 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 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 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 a 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) considerando os vários ambientes 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 considera os efeitos genotípicos como aleatórios (Resende & Thompson, 2004). Os modelos FAMM permitem inferências sobre valores genéticos, adaptabilidade, estabilidade e agrupamento de locais com base na interação genótipo x ambiente. 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 g 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 muito 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).

É importante relatar que o BLUP dos efeitos ge eliminam os chamados ruídos da interação genótipo x ambiente. 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

$Y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij} = \overline{Y}_{..} + (\overline{Y}_{i.} - \overline{Y}_{..}) + (\overline{Y}_{.j} - \overline{Y}_{..}) + \left(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}\right) + \varepsilon_{ij}$, 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

$I = \overline{Y}_{.j} + h_g^2(\overline{Y}_{i.} - \overline{Y}_{..}) + h_{ge}^2\left(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}\right)$ quando se considera os efeitos de ambiente como fixos (modelo misto) e por $I = \overline{Y}_{..} + h_g^2(\overline{Y}_{i.} - \overline{Y}_{..}) + h_e^2(\overline{Y}_{.j} - \overline{Y}_{..}) + h_{ge}^2\left(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}\right)$ 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 & 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.

G e L referem-se aos números de locais e de genótipos, respectivamente, e σ_ε^2 é a variância residual associada as médias Y_{ij}.

Verifica-se por estes índices que o BLUP de ge considera a herdabilidade dos efeitos da interação g x e, ou seja, elimina os ruíduos ou efeitos residuais, por ocasião do processo de predição de ge.

É importante relatar que, predito dessa forma, $\hat{g}_i$ fornece o efeito genotípico em um ambiente médio representado pelos L locais e $\hat{g}_i + g\hat{e}_{ij}$ fornece o efeito genotípico predito no local j ($\hat{g}_{ij}$). O efeito genotípico predito em um ambiente não avaliado (e com o 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. É importante relatar que os modelos BLUP que ajustam g e ge simultaneamente fornecem $\hat{g}_i^*$ e, $\hat{g}_{ij}$ pode ser obtido por $\hat{g}_i^* + g\hat{e}_{ij}^*$. O BLUP multivariado fornece $\hat{g}_{ij}$. A média desses, através dos locais, fornece $\hat{g}_i^*$ em qualquer dos três procedimentos mencionados. O BLUP univariado sem ajustar ge fornece $\hat{g}_i$.

Atualmente, procedimentos de interpretação mais simples têm sido preferidos 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 & Binns (1988) e modificações, tem sido enfatizados (Cruz & 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. 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; (iv) gera resultados na própria grandeza ou escala do caráter avaliado. Este último fator é bastante importante. Outros métodos como o de Lin & Binns fornecem resultados que não são interpretados diretamente como valores genéticos. O método de Annicchiarico depende, adicionalmente, de suposições de valores Z_{α}.

Na Tabela a seguir são apresentados alguns resultados comparativos envolvendo cinco clones (C) avaliados em 6 locais (L).

Tabela 1. 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 & Bins (Pi) e Annicchiarico (Annicch.). Avaliação de 5 clones (C) em 6 locais (L).

	L1	L2	L3	L4	L5	L6	VGMed	MHVG
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

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.

Verifica-se que a estatística MHPRVG produziu exatamente o mesmo ordenamento que as estatísticas de Lin & Binns (Pi) e de Annicchiarico e pode ser usada vantajosamente no contexto dos modelos mistos com efeitos genéticos aleatórios. A MHPRVG deve ser aplicada preferencialmente sobre os dados originais, expressando-os como ML/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.

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 & Bins e Annicchiarico. A MHPRVG pode ser aplicada via modelo multivariado ou via o modelo univariado do tipo g + ge com correção para heterogeneidade de variâncias. A MHPRVG é adequada para a seleção visando o plantio em ambientes com diferentes e variados padrões de interação g x e.

8. Análise Estatística de Múltiplas Características e Índices 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. Neste caso, o modelo multivariado é estruturado de forma a contemplar a covariância ambiental existente entre os caracteres. Assim, este modelo difere do modelo multivariado referido no tópico 7. 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.

9. Análise Estatística de Medidas Repetidas

No caso de medidas repetidas em cada indivíduo (ou tratamento) ao longo do tempo, várias alternativas (Resende, 2002a, p.522) existem para 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 genótipo x medições; (iii) modelo multivariado completo assumindo cada medida como um caráter diferente; (iv) modelo de regressão aleatória parcimônico como aproximação do modelo multivariado; (v) ajuste de splines cúbicas 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).

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. Assim, as opções mais interessantes aos melhoristas de plantas perenes são os modelos de repetibilidade (quando as suposições são atendidas), repetibilidade + interação genótipo x medições, ARH e SAD.

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 tópico 6, a aplicação do modelo de repetibilidade + interação genótipo x medição 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 a idade.

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 bloco x medição 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 medição.

$\sigma^2_{pp} = \sigma^2_{gb}$: 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 parcela x medição.

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 do ambiente da parcela x medição 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 medição. 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 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.

É importante relatar que 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 & Welham, 2003).

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 SAR), 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.

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, se deva 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.

10. Análise Estatística da Divergência Genética

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 ótimo.

Johnson & Wichern (1988) resumem as técnicas de análise multivariada da forma apresentada na sequência.

I. Métodos para Distinção entre Grupos
   (a) Análise de Agrupamento
   (b) Análise Discriminante

II. Métodos para Estudo da Estrutura de Covariância ou Correlação entre Variáveis
   (c) Componentes Principais
   (d) Análise de Fatores

Dentre estas 4 técnicas gerais, três (a, c e d) se relacionam mais diretamente com o estudo da divergência genética. 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.

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 & Regazzi, 1994; Cruz & 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 erros de predição 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.

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 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 relatado por Cruz & 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 valor fenotípico é dado por y = u + g + e, em que g é considerado efeito fixo. Na AF g é considerado aleatório e é desdobrado em $g = u + \Lambda f + \delta$, em que f denota o vetor aleatório dos escores fatoriais e 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 ambos, 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 genótipo x ambiente. 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 tópico 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.

11. Análise Biométrica da Seleção Recorrente Recíproca

A seleção recorrente recíproca (SRR) é a principal ferramenta para o melhoramento da média de cruzamentos interpopulacionais. Assim, deve ser utilizada no melhoramento de espécies em que a heterose é relevante nos caracteres de importância econômica. A SRR conduz ao melhoramento do híbrido interpopulacional. Para isto são melhoradas a heterose do cruzamento interpopulacional e pelo menos uma das populações envolvidas no cruzamento.

É importante relatar que tanto a seleção dos genitores a serem recombinados quanto a seleção dos indivíduos recombinados a serem cruzados devem ser baseadas nos efeitos aditivos interpopulacionais. No caso, a seleção dos referidos indivíduos recombinados (ou a serem recombinados, dependendo do esquema) deve ser baseada nos valores genéticos aditivos interpopulacionais preditos por $\hat{a}_{i(12)} = (1/2)\hat{a}_{g12} + h_c^2(y_{\text{intra}} - X\hat{\beta} - (1/2)\hat{a}_{g11} - W\hat{c}_{11} - S\hat{d}_{11})$, sob um modelo individual reduzido, em que $\hat{a}_{g12}$ é o valor genético aditivo interpopulacional do genitor da população 1 e $\hat{a}_{g11}$ é o valor genético aditivo intrapopulacional do genitor da população 1 obtidos sob um modelo reduzido bivariado. O componente $h_c^2 = \rho_a \sigma_{a_{1112}} / \sigma_{e11}^2$ é a herdabilidade correlacionada ou co-herdabilidade para a seleção dentro de genitor. O coeficiente ρ_a depende do tipo de progênie avaliada e equivale a (3/4) para meios irmãos. Verifica-se que a contribuição da seleção dentro de genitor é zero se a correlação entre as performances intra e inter for zero.

Com valores genéticos inter e intrapopulacionais preditos de forma independente, o índice (tendo como alvo o valor genético aditivo interpopulacional) para a seleção recorrente recíproca é dado por $I_{a_{12}} = a_{12} + (\sigma_{a12} / \sigma_{a11}) r_{G_{12}} a_{11}$ para a população 1 e $I_{a_{21}} = a_{21} + (\sigma_{a21} / \sigma_{a22}) r_{G_{21}} a_{22}$ para a população 2.

12. Análise Estatística 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 à abordagens bayesianas. Tais modelos estatísticos são misturas de distribuições, em que o número de densidades componentes é determinado pelo número de genótipos do QTL e 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 a 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 a 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:

$$
\begin{aligned}
y &= \mu + g + e \\
&= \mu + g_m + g_{nm} + e \\
&= \mu + \beta_L x_L + \beta_R x_R + g_{nm} + e \\
&= Xb + Z\beta_L x_L + Z\beta_R x_R + Zg_{nm} + \xi + \eta \text{ , em que:}
\end{aligned}
$$

$g = g_m + g_{nm}$: 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 FAMM.

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 e 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.

13. Referências Bibliográficas

ANNICCHIARICO, P. Cultivar adaptation and recomendation from alfalfa trials in Northern Italy. ***Journal of Genetics and Plant Breeding***, v. 46, p. 269-278, 1992.

CRESSIE, N.A.C. ***Statistics for spatial data analysis***. New York: Wiley, 1993. 900p.

CRUZ, C. D., CARNEIRO, P. C. S. ***Modelos biométricos aplicados ao melhoramento genétic*o. *Volume 2***. Vicosa, MG : Editora UFV, 2003. 585p.

CRUZ, C. D.; REGAZZI, O. J. ***Modelos biométricos aplicados ao melhoramento genético***. Viçosa: Universidade Federal de Viçosa, Imprensa Universitária, 1994. 390 p.

CULLIS, B.R.; GLEESON, A.C. Spatial analysis of field experiments-an extension at two dimensions. ***Biometrics***, v.47, p.1449-1460, 1991.

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

FEDERIZZI, L.C.; MILACH, S.C.K.; PACHECO, M.T. Melhoramento da Aveia. In: BORÉM, A. (Ed.) Melhoramento de Espécies Cultivadas. Viçosa: Editora UFV. 1999. p. 131-157.

FISHER, R. A. ***Statistical methods for research workers***. 1. ed. London: Oliver and Boyd, 1925. 314 p.

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.; CULLIS, B. R.; WELHAM, S.J.; THOMPSON, R. ***ASReml Reference manual.*** 2 Edition. Release 1.0. Biomathematics and Statistics Department - Rothamsted Research, Harpenden – England, 2002. 187 p.

GILMOUR, A.R.; CULLIS, B.R.; VERBYLA, A.P. Accounting for natural and extraneous variation in the analysis of field experiments. **J. Agric., Biol. Environ. Stat.**, v.2, p.269-293, 1997.

HILL, R.R.; ROSENBERGER, J.L. Methods for combining data from germplasm evaluation trials. ***Crop Science***, v.25, p.467-470, 1985.

JOHNSON, R. A.; WICHERN, D. W. ***Applied multivariate statistical analysis***. Englewood : Prentice Hall Inc., 1988. 594 p.

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.

MARTIN, R.J. The use of time-series models and methods in the analysis of agricultural field trials. **Commun. Stat. Theory Methods**, v.19, n.1, p.55-81, 1990.

LITELL, R.C. Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. ***J. Agric., Biol. Environ. Stat.,***v.7, n.4, p.472-491, 2002.

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.

RESENDE, M.D.V. ***Métodos estatísticos ótimos na análise de experimentos de campo***. Colombo: Embrapa Florestas. 2004. 65 p. (Documentos, 100).

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

RESENDE, M.D.V. ***Software Selegen–REML/BLUP***. Curitiba: Embrapa Florestas, 2002b. 67p (Documentos 77).

RESENDE, M. D. V. de. ; BARBOSA, M.H.P. Selection via simulated individual blup (blupis) based on family genotypic effects in sugarcane. ***Pesquisa Agropecuária Brasileira***, 2004. (submetido).

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***, Colombo, v. 28/29, p. 37-55, 1994.

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***, Colombo, n. 32/33, p. 18-45, 1996.

RESENDE, M.D.V.; STRINGER, J.K.; CULLIS, B.C; THOMPSON, R. Analysis of Interference and Environmental Trend in Field Trials by Joint Modelling of Competition and Spatial Variability. In: IUFRO Conference: Eucalyptus in a changing world. Aveiro: IUFRO, 2004. p. 330 – 332.

RESENDE, M.D.V.; THOMPSON, R.; WELHAM, S.J. Multivariate spatial statistical analysis in perennial crops. In: ***International Biometric Society Conference – British Region***, 2003. ***Proceedings.*** Reading: School of Applied Statistics – University of Reading. p.70-71.

RESENDE, M.D.V.; THOMPSON, R. Factor analytic multiplicative mixed models in the analysis of multiple experiments. ***Revista de Matemática e Estatística***, v.22, n.2, p. 31- 52, 2004.

SMITH, A.; CULLIS, B.R.; THOMPSON, R. Analysing variety by environment data using multiplicative mixed models and adjustment for spatial field trend. ***Biometrics***, v. 57, p. 1138-1147, 2001.

STRINGER, J.K.; CULLIS, B.R. Joint modelling of spatial variability and interplot competition. In McCOMB, J. A. (Ed). Proceedings of the 12th Australasian Plant Breeding Conference. Perth, Western Australia, 2002. p. 614-619.

STROUP, W.W.; MULITZE, D.K. Nearest neighbour adjusted best linear unbiased prediction. ***American Statistician***, v. 45, p. 194-200, 1991.

VENCOVSKY, R. Effective size of monoecious populations submitted to artificial selection. **Brazilian Journal of Genetics**, v.1, n.3, p.181-191, 1978.

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.