SEAGRO 2005
50ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBRAS)
11º Simpósio de Estatística Aplicada à Experimentação Agronômica (SEAGRO)
Apoio Financeiro: CAPES CNPq FAPESP FINEP FUNDAÇÃO ARAUCÁRIA

ESPECIFICAÇÃO DO MODELO DE ANÁLISE ESTATÍSTICA DE DADOS QUANTITATIVOS E SUAS IMPLICAÇÕES NA SELEÇÃO DE GENÓTIPOS EM PLANTAS

João Batista Duarte
Escola de Agronomia e Engenharia de Alimentos
Universidade Federal de Goiás (jbduarte@agro.ufg.br)

1 INTRODUÇÃO

A estatística representa, sem dúvida, uma poderosa ferramenta de apoio ao pesquisador para a análise dos dados por ele obtidos no processo de investigação científica. Alguns entusiastas dessa área do conhecimento chegam a afirmar que "a estatística permite extrair 'certezas' de incertezas", referindo-se, por exemplo, à estimação precisa de parâmetros desconhecidos por meio de um intervalo bastante estreito e de elevada confiança, a partir de amostras relativamente pequenas. O aproveitamento desse potencial, contudo, pode requerer o emprego de abordagens não convencionais, que necessitam ser melhor difundidas em certas áreas do conhecimento.

No melhoramento genético de plantas, por exemplo, ainda é comum o uso de análise baseada em modelo fixo para a estimação de médias de tratamentos (ex: genótipos), mesmo quando estes foram obtidos por amostragem numa população. Isto é, em situações em que o modelo é naturalmente misto, pois inclui, além de efeitos fixos, os efeitos aleatórios dos genótipos. Em boa parte dos casos, a modelagem mista é utilizada, com o rigor da suposição, apenas para a estimação de componentes de variância e para a construção de testes F apropriados na análise da variância.

Entre as razões que levam os melhoristas práticos a não utilizarem predições baseadas em modelos mistos estão a falta de vivência com estes métodos e a sua pequena divulgação (Bueno Filho, 1997). Acrescenta-se que os efeitos prejudiciais da abordagem mais tradicional, normalmente, são tidos como mínimos, a ponto de não recompensar os esforços com a adoção da nova abordagem. A ordem de classificação dos genótipos, em geral, não se altera no caso de ensaios que seguem delineamentos ortogonais e balanceados. Assim, na prática, a estimação de médias admitindo-se modelo fixo (ex: análise intrablocos em ensaios de competição de genótipos) quando, na verdade, é misto, não modificaria o resultado final da seleção.

Por outro lado, a ocorrência de desbalanceamento não planejado, decorrente da perda de parcelas, é um fato normal nesse tipo de experimentação. Ademais, nas fases preliminares do processo seletivo, quando os genótipos são numerosos e ainda possuem natureza aleatória (Piepho, 1994), é comum o uso de delineamentos como BIB (blocos incompletos balanceados), PBIB

(blocos incompletos parcialmente balanceados), blocos aumentados (blocos de Federer), que são naturalmente não-ortogonais (os dois últimos são, ainda, desbalanceados por construção). Nestes casos, a possibilidade de classificações genotípicas diferenciadas entre as duas abordagens é uma realidade e, como salienta Bueno Filho (1997), optar-se pela conveniência da suposição de um fator como fixo ou aleatório, pode estar longe de ser prática inofensiva.

Nos experimentos de avaliação de genótipos, em plantas, outro fato comum é a inclusão, entre os novos tratamentos genéticos em teste e com algum relacionamento de parentesco entre si, de cultivares comerciais como testemunhas. Nesse caso, um modelo matemático apropriado deve permitir, por exemplo, a acomodação de efeitos de tratamentos de duas naturezas: fixos para as testemunhas e aleatórios para os genótipos novos, objeto principal da avaliação. A aplicação rigorosa dessa suposição pode também implicar em resultados distintos daqueles de uma análise tradicional baseada em modelos fixos, com implicações no progresso genético obtido por seleção.

Outra característica do processo seletivo em melhoramento de plantas, nas suas fases preliminares, é a pequena quantidade de material de propagação para cada novo genótipo a ser avaliado. Isso limita drasticamente o uso de repetições para esses tratamentos genéticos, os quais são, com freqüência, avaliados numa só parcela experimental (sem repetições). Para lidar com esse tipo de limitação Federer (1956) propôs os delineamentos aumentados, os quais permitem ajustar as médias dos novos tratamentos para efeitos ambientais (blocos, linhas e/ou colunas) estimados a partir de testemunhas repetidas. O autor apresentou também os métodos de análise estatística desses delineamentos, baseados em *quadrados mínimos ordinários* (OLS), assentados, portanto, na suposição de independência entre observações.

A pouca disponibilidade de material (sementes, tubérculos etc.), entretanto, força também o melhorista a adotar unidades experimentais de pequeno tamanho (parcelas curtas e estreitas), usualmente com apenas uma ou duas fileiras de plantas. E isso, sabidamente, aumenta a possibilidade de violação da independência entre observações assumida pelo método OLS, haja vista a maior probabilidade de correlação entre observações de parcelas vizinhas. Este fenômeno, denominado *autocorrelação espacial*, pode comprometer seriamente a comparação de tratamentos. Es & Es (1993) demonstraram que, sob esse tipo de correlação, os testes estatísticos associados a contrastes de tratamentos cujas parcelas estiveram separadas por pequenas distâncias têm maior probabilidade de erro tipo II; enquanto os contrastes de tratamentos cujas parcelas estiveram separadas por distâncias maiores foram testados com maior probabilidade de erro tipo I. A solução desse problema passa pelo enfoque da análise estatística de dados espacialmente correlacionadas, sob *quadrados mínimos generalizados* (*GLS*), já amplamente descrita (Besag & Kempton, 1986; Gleeson & Cullis, 1987; Cullis *et al.*, 1989; Stroup & Mulitze, 1991; Grondona & Cressie, 1991;

Zimmerman & Harville, 1991; Stroup *et al.*, 1994; entre outros). Apesar da vasta literatura correlata, na prática, esses métodos estão longe de serem realmente incorporados como rotina nos programas de melhoramento de plantas.

O objetivo deste trabalho é ilustrar possíveis implicações da adoção de abordagens analíticas simplificadas, sobre a seleção de genótipos de plantas experimentalmente testados, quando a natureza dos dados não satisfaz as suposições que apóiam tais abordagens. Para isso, foram considerados três casos, extraídos de Duarte (2000), enfocando: *i)* a estimação e a predição sob modelo linear misto, com ênfase na ordenação das médias de tratamentos genéticos; *ii)* o efeito da recuperação de informação intergenotípica dos novos genótipos em teste (de efeitos aleatórios) sobre a seleção daqueles superiores às cultivares testemunhas (de efeitos fixos); e *iii)* a influência da adoção de uma abordagem de análise estatística espacial na seleção de genótipos, quando as observações não forem espacialmente independentes.

2 ESTIMAÇÃO E PREDIÇÃO SOB MODELO LINEAR MISTO COM ÊNFASE NA ORDENAÇÃO DAS MÉDIAS GENOTÍPICAS

Nesta seção procura-se ilustrar, por meio de dois exemplos simulados, as possíveis diferenças na classificação das médias genotípicas, quando estas forem obtidas por modelos fixo ou misto, de análise estatística. O desenvolvimento centra-se na abordagem de modelos lineares mistos, tomando-se como exemplo, sem perda de generalidade, um modelo de delineamento experimental em blocos (detalhes em Duarte & Vencovsky, 2001).

Consideraram-se, assim, a tratamentos genéticos (genótipos) de efeitos g_i ($i=1,2,...,a$) e b blocos (completos ou incompletos) de efeitos b_j ($j=1,2,...,b$). Com o propósito de generalização, faz-se n_{ij} ser o número de vezes que o tratamento i aparece no bloco j ($n_{ij}=0,1,2,...$). Portanto: $\Sigma_i\Sigma_j n_{ij} = n$ (número de observações); $\Sigma_i n_{ij}=n_{.j}=k_j$ (tamanho ou número de parcelas do bloco j); e $\Sigma_j n_{ij}=n_{i.}=n_i$ (número de repetições do tratamento i); além de que: $\Sigma_i n_i=\Sigma_j k_j=n$. Denota-se ainda por Y_{ijr} a observação num caráter ou variável aleatória Y (observável), relativa à *r-ésima* parcela ($r=1,2,...,n_i$) que recebeu o tratamento i, identificada também pelo bloco j. Um modelo linear que caracteriza esse conjunto de dados pode ser escrito como:

$$Y_{ijr} = m + b_j + g_i + e_{ijr}\ ; \qquad \text{com:}\quad g_i \sim N(0,\ \sigma_g^2);\quad e_{ijr} \sim N(0, \sigma_e^2);$$
$$E(Y_{ijr}) = m + b_j\ ;\ \text{e}\quad Var(Y_{ijk}) = \sigma_g^2 + \sigma_e^2\,.$$

Neste modelo, o efeito de bloco (b_j) foi assumido como fixo e o de tratamento (g_i), como aleatório. A constante m é de natureza sempre fixa e e_{ijr} é uma variável aleatória não observável. Isso caracteriza o que se conhece na literatura por *modelo linear misto*, pois incorpora uma mistura de tipos de efeitos, fixos e aleatórios (Searle, 1987). Dessa forma, os tratamentos testados representam uma amostra de uma população de genótipos, cujas respostas são distribuídas *normalmente*, em torno de uma média comum ($\mu_P = m + \overline{b}$) e com variância σ_g^2; ou seja, os tratamentos são realizações de variáveis aleatórias não observáveis, as quais correspondem aos efeitos g_i (desvios genotípicos aleatórios em relação à média μ_P). O tratamento estatístico desse tipo de modelo, no campo do melhoramento genético vegetal, tem recebido a denominação de *análise com recuperação da informação intervarietal* ou *intergenotípica* (Wolfinger *et al*.,1997; Federer, 1998).

Matricialmente, a expressão que generaliza essa e outras modelagens mistas alternativas pode ser escrita a partir do vetor $\mathbf{y}_{(nx1)}$ de observações, na forma do chamado *modelo linear misto geral* (Henderson, 1984):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon \;\; ; \quad \text{com:} \quad \varepsilon \sim N(\phi,\mathbf{R}); \quad \gamma \sim N(\phi,\mathbf{G});$$
$$E(\mathbf{y}) = \mathbf{X}\beta; \;\; \text{e} \;\; Var(\mathbf{y}) = \mathbf{V}_{(n)} = \mathbf{ZGZ'} + \mathbf{R} \, .$$

Neste caso, tem-se: todos os efeitos fixos reunidos no vetor paramétrico $\beta_{(px1)}$; os efeitos aleatórios no vetor paramétrico $\gamma_{(qx1)}$, exceto os erros que compõem o vetor $\varepsilon_{(nx1)}$; $\mathbf{X}_{(nxp)}$ e $\mathbf{Z}_{(nxq)}$ representam as matrizes de incidências dos efeitos contidos em β e γ, respectivamente; $\mathbf{G}_{(q)}$ e $\mathbf{R}_{(n)}$ são as matrizes de variâncias-covariâncias dos vetores aleatórios γ e ε, respectivamente; e as covariâncias entre vetores diferentes são assumidas nulas (Henderson, 1984). Aqui, por simplificação, adotar-se-á: $\mathbf{G} = \mathbf{I}_{(a)}\sigma_g^2$ e $\mathbf{R} = \mathbf{I}_{(n)}\sigma_e^2$, em que $\mathbf{I}_{(.)}$ denota uma matriz identidade e $a=q$ (número de níveis do fator aleatório).

A estimação dos efeitos fixos no modelo (BLUE) e suas funções de interesse, bem como a predição dos efeitos aleatórios (EBLUP), foram obtidas simultaneamente por meio das chamadas *equações de modelo misto* (EMM), propostas por Henderson, em 1948 (Littell *et al*., 1996; Henderson, 1984):

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \widetilde{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

Para isso, utilizaram-se as estimativas $\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$, obtidas via procedimento de *máxima verossimilhança restrita* (REML), em substituição às matrizes $\mathbf{G}$ e $\mathbf{R}$.

Para melhor entender as conseqüências da suposição de aleatoriedade dos efeitos de tratamentos sobre suas estimativas de médias, é importante explicitar, como ilustram Searle *et al.* (1992), a expressão do preditor do efeito genotípico, ou seja, o *BLUP*(g_i):

$$BLUP(g_i) = \widetilde{g}_i = \frac{n_i\sigma_g^2}{\sigma_e^2 + n_i\sigma_g^2}(\overline{Y}_i - \mu^0)$$

O estimador da média genotípica deve agregar a $\widetilde{g}_i$ uma estimativa do efeito ambiental médio ($\mu_P = m + \overline{b}$); representando, portanto, uma função linear de parâmetros fixos e aleatórios, $\mathbf{w} = \mathbf{L}'\beta + \gamma$. Seu estimador, com propriedades de preditor linear não viesado (Searle *et al.*, 1992), é: *BLUP*($\mathbf{w}$) = $\widetilde{\mathbf{w}} = \mathbf{L}'\beta^0 + \widetilde{\gamma}$; em que a matriz $\mathbf{L}'_{[ax(1+b)]}$, neste caso, possui suas linhas todas iguais a [*1* k_1/n k_2/n ... k_b/n], o que gera uma média ponderada dos blocos por seus respectivos tamanhos ($m^0 + \overline{b}^0$). Logo, a média *BLUP* do genótipo *i* pode ser escrita como: $BLUP(\mathbf{w}) = BLUP(\mu_P + g_i) = \hat{\mu}_P + \widetilde{g}_i = (m^0 + \overline{b}^0) + \widetilde{g}_i$.

Dois exemplos foram gerados por simulação. No primeiro deles, com resultados expressos na Figura 1, consideraram-se dez tratamentos genéticos (genótipos) de efeitos aleatórios $g_i \sim N(0, \sigma_g^2 = 0{,}25)$ e de uma só população. No segundo (Figura 2), foram consideradas três populações de efeitos fixos (P1, P2 e P3), cada uma com quatro genótipos de efeitos aleatórios dentro de populações: $g_i \sim N(1, \sigma_g^2 = 0{,}15)$ se $i \in$ P1; $g_i \sim N(2, \sigma_g^2 = 0{,}05)$ se $i \in$ P2; e $g_i \sim N(3, \sigma_g^2 = 0{,}2)$ se $i \in$ P3. Nos dois casos, assumiram-se ainda: $\mu = 10$; b_j gerados sob $N(0, S_b^2 = 0{,}2)$ e $e_{ijr} \sim N(0, \sigma_e^2 = 2{,}00)$.

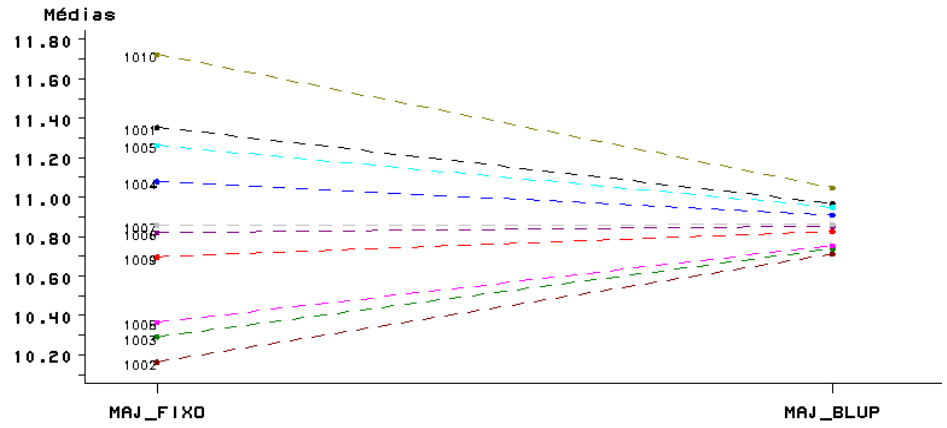

Figura 1. Efeito *shrinkage* sobre médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP). Os números (1001 a 1010) identificam os genótipos, num ensaio simulado de blocos completos casualizados com: $\mu = 10$; $b_j \sim N(0, S_b^2 = 0{,}20)$; $g_i \sim N(0, \sigma_g^2 = 0{,}25)$; e $e_{ijr} \sim N(0, \sigma_e^2 = 2{,}00)$.

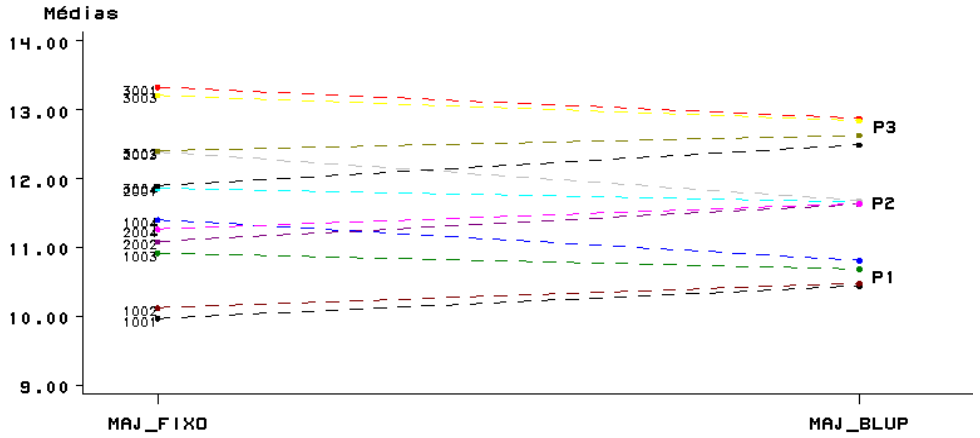

Figura 2. Ordenamento de médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP), para um conjunto de tratamentos oriundos de três populações (P1, P2 e P3). Dados simulados para um ensaio em blocos ao acaso sob: $\mu=10$; $b_j \sim N(0, S_b^2=0,2)$; $e_{ijr} \sim N(0, \sigma_e^2=2,0)$; e $g_i \sim N(1, \sigma_g^2=0,15)$ se $i \in$ P1; $g_i \sim N(2, \sigma_g^2=0,05)$ se $i \in$ P2; e $g_i \sim N(3, \sigma_g^2=0,2)$ se $i \in$ P3.

A análise teórica desses exemplos permitiu concluir: 1) a abordagem de modelos mistos com tratamentos aleatórios, em geral, produz médias mais uniformes (efeito *shrinkage*) para os tratamentos do que a análise intrablocos; 2) a metodologia de modelos mistos implica em predições e ordenações genotípicas notadamente diferentes em relação às análises tradicionais (médias marginais e análise intrablocos), quando a variabilidade genética relativa (σ_g^2/σ_e^2) for baixa e os experimentos forem não-ortogonais e desbalanceados; 3) se os tratamentos genéticos forem oriundos de populações diferentes, o fato de a predição *BLUP* levar em conta as variâncias genotípicas de cada população, ou melhor a herdabilidade associada, pode determinar diferentes classificações dos tratamentos em relação a uma análise intrablocos, mesmo sob ortogonalidade e balanceamento.

Essas constatações permitem concluir, ainda, que é um equívoco admitir que na análise de um modelo com um fator aleatório, ao invés de fixo, apenas os componentes de variância (esperanças de quadrados médios) e os teste F podem se alterar.

3 EFEITO DA RECUPERAÇÃO DE INFORMAÇÃO INTERGENOTÍPICA NA SELEÇÃO DE GENÓTIPOS SUPERIOES À TESTEMUNHA

Nesta seção, ilustra-se, a partir de dados reais, as implicações de se assumir efeitos genotípicos como aleatórios (com recuperação de informação intergenotípica) ou fixos, sobre a seleção de genótipos superiores à(s) cultivar(es) testemunha(s). Os dados referem-se a um ensaio de avaliação de linhagens de soja, conduzido na localidade Areão, município de Piracicaba-SP, em 1992/1993 (Programa de Melhoramento da Soja, Setor de Genética Aplicada às Espécies Autógamas, Departamento de Genética da ESALQ/USP). Avaliaram-se 1260 progênies, distribuídas em 28 blocos de tamanho k=55 parcelas (45 progênies + 10 testemunhas). Em decorrência de observações perdidas ou discrepantes, das 1540 parcelas originais foram aproveitadas 1379 observações para o caráter produtividade, perfazendo a avaliação de 1228 linhagens experimentais.

A partir de um modelo linear geral para delineamento em blocos ($Y_{ij}= \mu + \beta_j + \tau_i + \varepsilon_{ij}$), as análises implementadas (modelagens alternativas) podem ser sumarizadas em quatro tipos:

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

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

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

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

Para fins de ilustração, tomou-se apenas uma amostra aleatória de dez das 1228 progênies, mais as duas testemunhas melhor classificadas. Algumas relações entre as médias genotípicas ajustadas por esses modelos, podem ser visualizadas nas Figuras 3 a 6.

De uma maneira geral, pôde-se constatar: 1) a possibilidade de troca de posições relativas das linhagens entre os diferentes modelos (diferentes linhagens liderando a classificação); 2) a tendência dos *Modelos 3* e *4* produzirem médias mais uniformemente distribuídas do que os outros

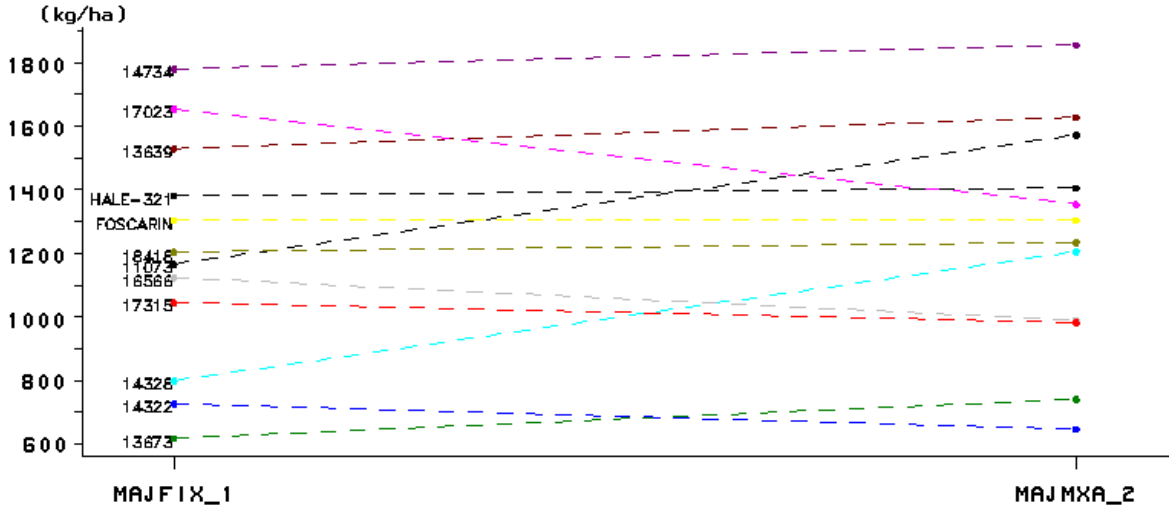

Figura 3. Relações entre médias genotípicas ajustadas pela análise intrablocos (MAJFIX_1) e pela análise com recuperação da informação interblocos (MAJMXA_2), para uma amostra de dez linhagens experimentais e duas testemunhas (HALE-321 e FOSCARIN), entre *1228* genótipos avaliados em blocos aumentados, no ensaio AREÃO-12 (grupo Precoce, ano de 1992/93, ESALQ-USP).

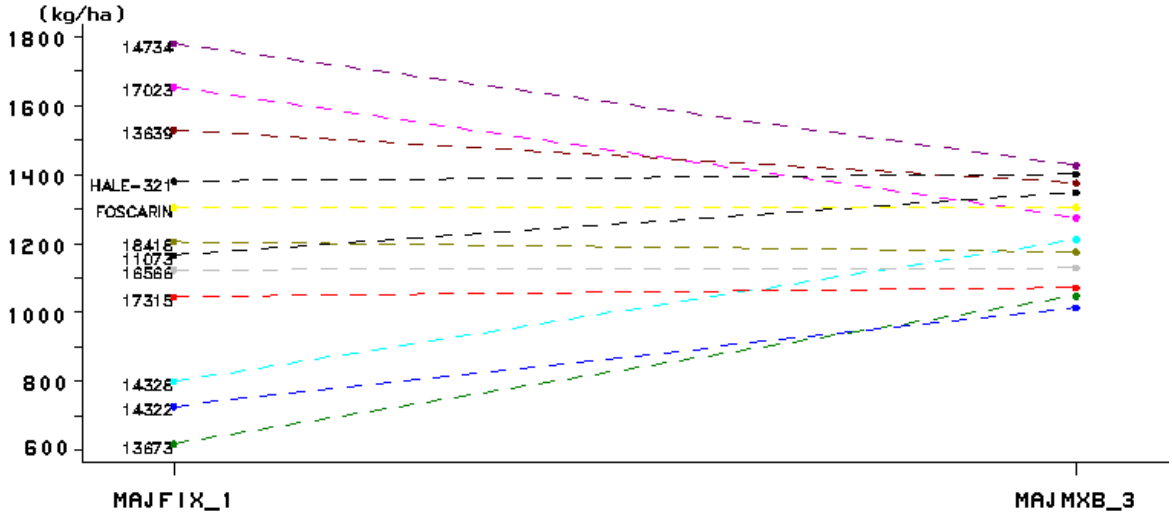

Figura 4. Relações entre médias genotípicas ajustadas pela análise intrablocos (MAJFIX_1) e pela análise com recuperação da informação intergenotípica (MAJMXB_3), para uma amostra de dez linhagens experimentais e duas testemunhas (HALE-321 e FOSCARIN), entre *1228* genótipos avaliados em blocos aumentados, no ensaio AREÃO-12 (grupo Precoce, ano de 1992/93, ESALQ-USP).

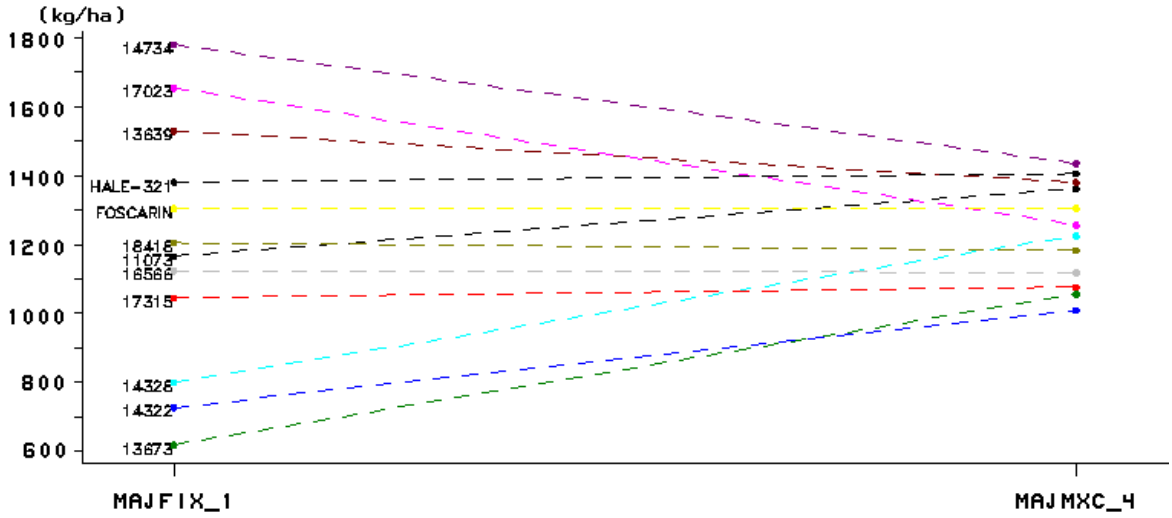

Figura 5. Relações entre médias genotípicas ajustadas pela análise intrablocos (MAJFIX_1) e pela análise recuperando as informações interblocos e intergenotípica (MAJMXC_4), para uma amostra de dez linhagens experimentais e duas testemunhas (HALE-321 e FOSCARIN), entre *1228* genótipos avaliados em blocos aumentados, no ensaio AREÃO-12 (grupo Precoce, 1992/93, ESALQ-USP).

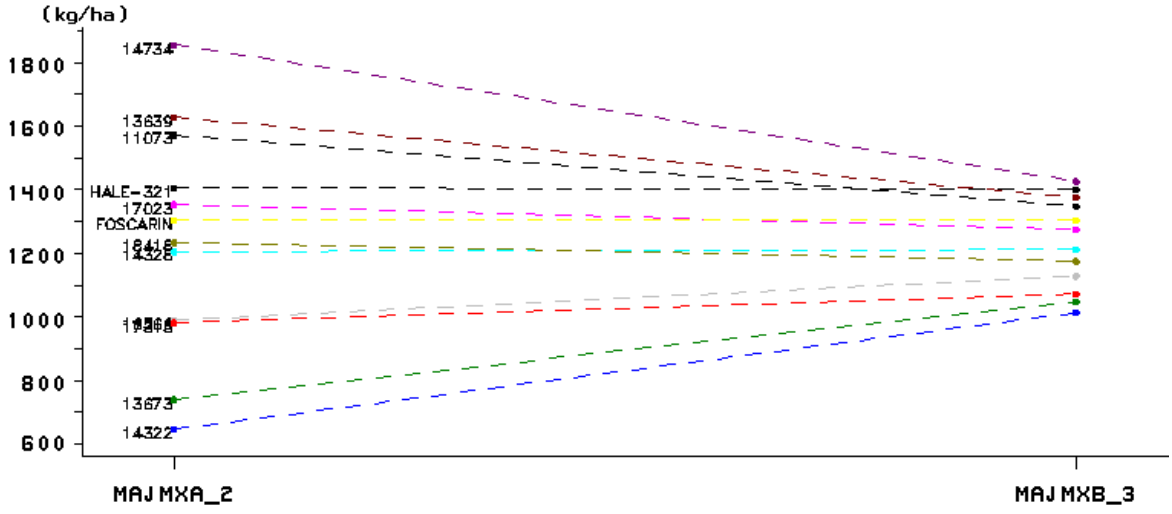

Figura 6. Relações entre médias genotípicas ajustadas pelas análises com recuperação da informação interblocos (MAJMXA_2) e da informação intergenotípica (MAJMXB_3), para uma amostra de dez linhagens experimentais e duas testemunhas (HALE-321 e FOSCARIN), entre *1228* genótipos avaliados em blocos aumentados, no ensaio AREÃO-12 (grupo Precoce, 1992/93, ESALQ-USP).

modelos (efeito *shrinkage* sobre as médias de progênies, não sobre as de testemunhas); e 3) os diferentes números de linhagens superando as melhores testemunhas, sobretudo quando se recupera (*Modelos 3* e *4*) ou não (*Modelos 1* e *2*) informação intergenotípica.

Em síntese, essas avaliações permitem concluir que o uso de diferentes modelos para a obtenção das médias genotípicas pode levar a seleções diferenciadas. E, a despeito de alguma concordância nos ordenamentos genotípicos desses modelos, o fato é que genótipos bem classificados por alguns deles podem ser descartados por outros. Considerando-se que a expectativa normal de um ciclo seletivo, numa espécie já bastante melhorada como a soja, é a liberação de uma ou duas cultivares, essa diferenciação pode determinar o sucesso ou fracasso do programa.

4 ANÁLISE ESTATÍSTICA ESPACIAL E SELEÇÃO DE GENÓTIPOS EM PROGRAMAS DE MELHORAMENTO DE PLANTAS

Nesta seção busca-se demonstrar os benefícios potenciais de uma abordagem de análise estatística espacial, na seleção de genótipos superiores num programa de melhoramento. Conforme já comentado, tais benefícios são evidenciados comparativamente a uma análise estatística clássica, sobretudo quando as observações experimentais não forem espacialmente independentes. Maiores detalhes acerca do tema e do caso aqui estudado estão em Duarte & Vencovsky (2005).

O material experimental que fundamenta a presente ilustração consistiu de um ensaio de competição de linhagens de soja, também conduzido no local Areão, município de Piracicaba-SP, em 1994/1995 (Programa de Melhoramento da Soja, ESALQ/USP). O delineamento experimental foi em blocos aumentados, com $t=5$ cultivares testemunhas e $p=110$ novas linhagens, distribuídos em $b=4$ blocos de aproximadamente 50 parcelas. A parcela correspondeu apenas a duas fileiras de plantas. Somente os dados de produtividade de grãos (kg/ha) foram considerados. Para implementar uma análise estatística espacial são necessárias, também, as distâncias (m) correspondentes às coordenadas geográficas do centro de cada parcela, na grade de campo do experimento.

Os dados foram submetidos às análises estatísticas por dois modelos matemáticos: *i)* modelo assumindo observações espacialmente independentes; e *ii)* modelo admitindo correlação espacial entre observações. Este último foi ajustado conforme a proposta de Zimmerman & Harville (1991), denominada "modelo linear de campo aleatório" (*random field linear model – RFLM*). Nos dois casos, os efeitos das novas linhagens foram admitidos como aleatórios e relacionados a uma só população. Logo, ambos são modelos mistos, com diferença apenas na suposição associada ao efeito do erro experimental.

Esses modelos acomodaram, ainda, efeitos de tratamentos de duas naturezas: fixos para as testemunhas (cinco populações individualizadas, sem variação dentro) e aleatórios para as linhagens dentro da sexta população (esta também supostamente fixa, mas com variação dentro). Assim, as observações foram caracterizadas como:

$$Y_{ijk} = \mu + b_j + c_k + g_{i(k)} + e_{ijk}$$

em que:

Y_{ijk}: é a observação na parcela que recebeu o genótipo *i* relacionado à população *k*, no bloco *j*;
μ : é a constante comum a todas as observações;
b_j : é o efeito fixo do *j-ésimo* bloco *(j=1,2,...,b)*;
c_k : é o efeito fixo da *k-ésima* população *(k=1, 2, ..., t, t+1)*;
$g_{i(k)}$: é o efeito do *i-ésimo* genótipo relacionado à *k-ésima* população, assumido fixo e nulo se o genótipo for uma testemunha *(i=1)*, ou aleatório com distribuição $N(0, \sigma_g^2)$ independente, se o genótipo for uma nova linhagem *(i=1,2,...,p)*; e
e_{ijk}: é o erro experimental aleatório associado à *ijk-ésima* parcela, assumido independente (covariância nula entre erros de parcelas diferentes) e com distribuição $N(0, \sigma_e^2)$, na primeira análise (*i*); e, na segunda (*ii*), $e_{ijk} \sim N\,[0, C(h)]$; sendo *C(h)* a (co)variância entre dois erros de parcelas separadas por uma distância *h*, com $h \geq 0$ (esses erros são aqui denotados por $e_{(s)}$ e $e_{(s+h)}$, em que *s* indica a posição espacial da *ijk-ésima* parcela).

Ademais, na abordagem *RFLM*, *C(h)* é definida como (Littell *et al*., 1996):

$$C(h) = \begin{cases} \sigma^2 \text{ , se } h = 0 \text{ ; e} \\ \sigma_{e_{(s)}, e_{(s+h)}} = \sigma^2[f(h)] \text{ , se } h > 0 \end{cases}$$

Portanto, a covariância dos erros é assumida como uma função da distância que separa as correspondentes parcelas (*f(h)*). Esta, entretanto, não é preestabelecida, mas sim, estimada do "ensaio de uniformidade" sugerido pelos resíduos do ajuste do modelo com erros independentes, $\hat{e}_{ijk}$. Na abordagem de *RFLM,* essa função é estimada por procedimento de geoestatística via ajuste de *variograma* ou *semivariograma*.

Expressando-se as observações por um vetor **y**, ambos os modelos podem ser representados vetorialmente pelo *modelo linear misto geral* $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$, já descrito. Os efeitos genotípicos aleatórios (γ) foram assumidos, sem perda de generalidade do modelo, terem distribuição normal com média nula (ϕ) e matriz de (co)variâncias $\mathbf{G}=\mathbf{I}\,\sigma_g^2$; e os erros experimentais, com distribuição normal de média nula e matriz genérica de (co)variâncias **R**. Assim, no primeiro modelo (*i*) tem-se que $\mathbf{R}=\mathbf{I}\,\sigma_e^2$; enquanto no outro (*ii*): **R**=Σ (matriz não-diagonal e de estrutura definida pela

função geral de covariância e pelo alcance da autocorrelação espacial). Os componentes de variância σ_g^2 e σ_e^2 foram estimados por *REML*.

Os resíduos do ajuste do modelo sob independência ($\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$) foram, então, utilizados para estimar os parâmetros da estrutura de correlação espacial via semivariograma. Esses parâmetros especificam a matriz **R**=Σ, que é utilizada na obtenção de estimativas, predições genotípicas e testes estatísticos, livres dos efeitos da autocorrelação estimada. Essa etapa é processada através das *equações do modelo misto*:

$$\begin{bmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{X} & \mathbf{X}'\Sigma^{-1}\mathbf{Z} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{X} & \mathbf{Z}'\Sigma^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{y} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{y} \end{bmatrix}$$, cujas soluções de *GLS* são conhecidas.

Para avaliar tão somente os efeitos do ajuste espacial sobre os resultados da análise estatística utilizou-se, neste ajuste, o mesmo valor da estimativa de σ_g^2 obtido na análise sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$.

A presença da autocorrelação espacial, no presente caso, pode ser atestada nas Figuras 7 e 8. O fato é um indicativo de violação da independência espacial entre observações, assumida pelo primeiro modelo de análise (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$). Observa-se que os resíduos não se distribuíram mesmo de forma aleatória no campo experimental, mas, pelo contrário, houve uma tendência nítida de os maiores valores $\hat{e}_{ijk}$ concentrarem-se na parte posterior do mapa de campo (Figura 7). Além disso, a configuração dos pontos no variograma (Figura 8) é típica dos processos estocásticos com dependência espacial; isto é, com variabilidade decrescente à medida que a distância diminui. Evidencia-se que a partir de 20 m (alcance da correlação espacial), a variabilidade tende a se estabilizar um pouco acima de 120.000 (kg/ha)^2. A função contínua exata que se ajustou aos pontos foi o modelo exponencial de semivariâncias, definido por: $S(h) = \sigma^2[1 - exp(\frac{-3h}{a})]$, com σ^2=126450 (kg/ha)^2 e a = 20,4 m, no presente caso. Por conseguinte, a respectiva função de autocovariância ficou dada por $C(h) = 126450\, exp(\frac{-h}{6,8})$, que, por sua vez, definiu a matriz de (co)variâncias residuais $\mathbf{R}_{(n)}$=Σ, cujos elementos da diagonal principal foram iguais a 126450 e os fora desta diagonal, $126450\, exp(\frac{-h}{6,8})$, sendo h a distância que separa cada duas parcelas identificadas por uma linha e uma coluna da matriz.

As vantagens estatísticas da abordagem espacial apareceram, principalmente, na capacidade de discriminação dos genótipos (Tabela 1). Observa-se que a variação entre as seis populações fixas não foi significativa na primeira análise (5% de probabilidade), mas atingiu elevada signifi-

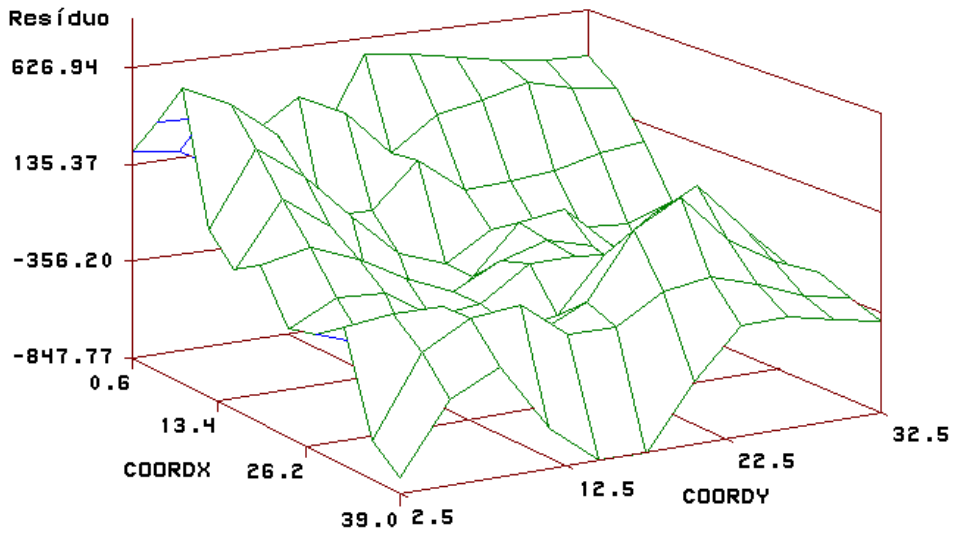

Figura 7. Resíduos (kg/ha) do modelo de blocos aumentados com recuperação de informação entre novos tratamentos, sob erros independentes, tomados em função das coordenadas, em metros, dos centros das respectivas parcelas (**COORDX** e **COORDY**).

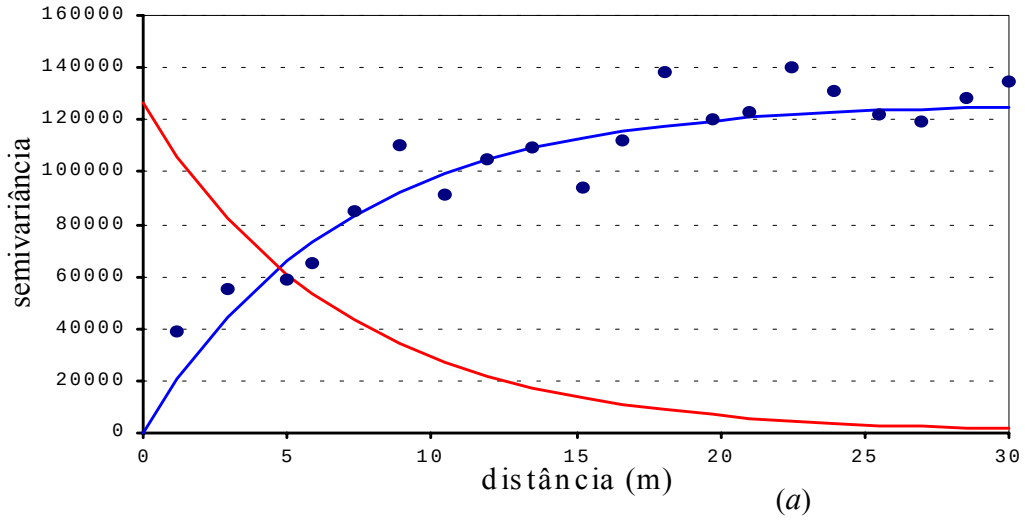

Figura 8. Variograma amostral (pontos) dos resíduos de uma análise de blocos aumentados assumindo erros independentes e seu respectivo ajuste (linha azul) pelo modelo exponencial de semivariâncias (a=20,4m; σ^2=126450kg^2/ha^2) (a linha vermelha ilustra a correspondente função de autocovariância espacial).

ficância estatística (p<0,01) na análise espacial. No desdobramento dessa fonte de variação, notou-se uma elevação nos valores de F, também em favor da análise espacial, tanto na detecção de diferenças entre as testemunhas, como na diferenciação entre as testemunhas e as novas linhagens. Nos três contrastes escolhidos para ilustrar a comparação entre linhagens novas, a vantagem da análise espacial foi ainda maior. Enquanto a análise sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$ não captou diferença alguma entre esses genótipos (p>0,95), a análise espacial apontou dois dos três contrastes como significativos (p<0,025).

Tabela 1. Testes de alguns efeitos genotípicos obtidos a partir dos modelos de análise espacial e não espacial (dados de produtividade de grãos, em kg/ha, num ensaio de competição de linhagens de soja delineado em blocos aumentados – Areão: 1994/95, ESALQ-USP).

Fontes de Variação	NDF*	Análise não espacial			Análise espacial		
		DDF*	F	Pr > F	DDF*	F	Pr > F
Populações	5	12,0	2,71	0,0727	31,9	3,95	0,0067
Testemunhas (T)	4	11,4	0,93	0,4793	31,5	1,79	0,1547
T vs. Novas Linhagens	1	15,5	9,89	0,0065	33,3	10,28	0,0030
Linhag.G1 vs Linhag.G3	1	0,20	0,00	0,9859	23,7	0,46	0,5064
Linhag.G1 vs Linhag.G24	1	0,20	0,01	0,9712	25,8	6,00	0,0214
Linhag.G3 vs Linhag.G24	1	0,20	0,01	0,9575	25,0	9,82	0,0044

* - NDF e DDF são, respectivamente, os números de graus de liberdade do numerador e do denominador da estatística F (os últimos obtidos pela aproximação de Satterthwaite).

A maior capacidade de diferenciação dos novos genótipos, em favor da análise espacial, foi também confirmada pelos valores preditos (EBLUP) dos efeitos genotípicos. Enquanto na primeira análise estes variaram entre -98,2 kg/ha e 100,5 kg/ha (dados completos em Duarte, 2000), na análise espacial essa amplitude foi superior a 500 kg/ha (de –337 kg/ha a 200,5 kg/ha). Os menores erros padrão a eles associados também confirmaram a melhor discriminação das linhagens pelo segundo modelo de análise (Tabela 2).

Considerando-se uma seleção hipotética de 25% das linhagens mais produtivas (28 em 110 genótipos), houve uma coincidência de apenas 46% entre as linhagens que seriam selecionadas nas duas análises (Tabela 2). Ademais, entre os genótipos selecionados na análise mais tradicional, pelo menos 30% ocupariam más posições de classificação (acima da 50ª) na análise espacial (ex.: USP 93-2048, USP 93-2393, USP 93-2153 e USP 93-2198). Em contrapartida, quatro linhagens classificadas entre as dez mais produtivas na análise espacial seriam descartadas na análise alternativa (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$). Reiterando a informação de que, no final de um ciclo seletivo em espécies vegetais cultivadas, dificilmente são liberadas aos produtores mais que duas cultivares, esse descarte pode determinar o insucesso do programa de melhoramento.

Tabela 2. Valores preditos dos efeitos genotípicos individuais (*EBLUP*) de 28 linhagens de soja, erros padrão associados e respectivo ordenamento segundo dois modelos de análise estatística (dados de produtividade de grãos, em kg/ha, num ensaio de competição de 110 linhagens e 5 cultivares testemunhas, em blocos aumentados – Areão/1994-95, ESALQ/USP).

Genótipo	Modelo de análise sob $R=I\,\sigma_e^2$			Modelo de análise espacial ($R=\Sigma$)		
	EBLUP	Erro padrão	Ordem	EBLUP	Erro padrão	Ordem
USP 93-2802	100,48	123,96	1	200,50	103,71	1
USP 93-2850	96,97	123,98	2	172,97	103,68	2
USP 93-2547	76,77	123,98	3	66,91	103,27	20
USP 93-2075	76,37	123,95	4	64,51	107,80	22
USP 93-2302	72,64	123,96	5	107,14	103,34	11
USP 93-2114	70,08	123,95	6	42,59	107,77	31
USP 93-2623	69,27	123,96	7	67,11	103,33	19
USP 93-2642	68,46	123,98	8	13,76	111,13	50
USP 93-2722	66,80	123,96	9	83,48	103,35	14
USP 93-2753	65,68	123,96	10	156,46	103,34	4
USP 93-2171	62,45	123,95	11	126,62	103,54	6
USP 93-2159	58,72	123,96	12	150,22	103,69	5
USP 93-2479	56,78	123,98	13	-21,24	111,08	65
USP 93-2881	53,56	123,96	14	172,84	103,34	3
USP 93-2187	53,41	124,01	15	47,13	111,11	29
USP 93-2037	52,79	123,95	16	5,82	111,74	53
USP 93-2148	52,43	123,96	17	38,83	103,63	35
USP 93-2475	50,86	123,96	18	33,96	103,35	37
USP 93-2474	48,70	124,01	19	64,17	111,68	23
USP 93-2048	48,53	123,95	20	-45,25	104,17	83
USP 93-2198	47,49	123,96	21	-32,91	111,20	75
USP 93-2266	46,90	124,01	22	25,62	103,35	42
USP 93-2153	46,51	123,95	23	-40,82	103,73	81
USP 93-2916	44,66	124,01	24	17,93	103,33	48
USP 93-2418	43,98	124,01	25	76,09	102,76	18
USP 93-2393	40,76	123,96	26	-43,15	103,34	82
USP 93-2077	38,15	123,98	27	-4,27	107,06	55
USP 93-2985	37,30	123,95	28	25,77	108,87	41

Uma inspeção no mapa de campo acerca da localização das parcelas onde foram alocadas as linhagens selecionadas nas duas abordagens revela outra surpresa (Figura 9). Quando o modelo não-espacial foi utilizado, os genótipos selecionados vieram exclusivamente da faixa lateral esquerda do campo experimental, provavelmente a sua área mais fértil. Porém, quando o ajuste espacial foi levado em conta, os genótipos selecionados vieram de parcelas espalhadas numa maior extensão da área experimental, o que revela uma maior independência desta análise em relação aos efeitos da variação local. Enfim, considerando-se que a divergência nas duas seleções decorreu basicamente do ajustamento genotípico para efeitos de posição, de natureza puramente ambiental,

conclui-se que, em situações similares, o uso da análise espacial pode garantir uma maior eficiência à seleção no contexto de um programa de melhoramento genético.

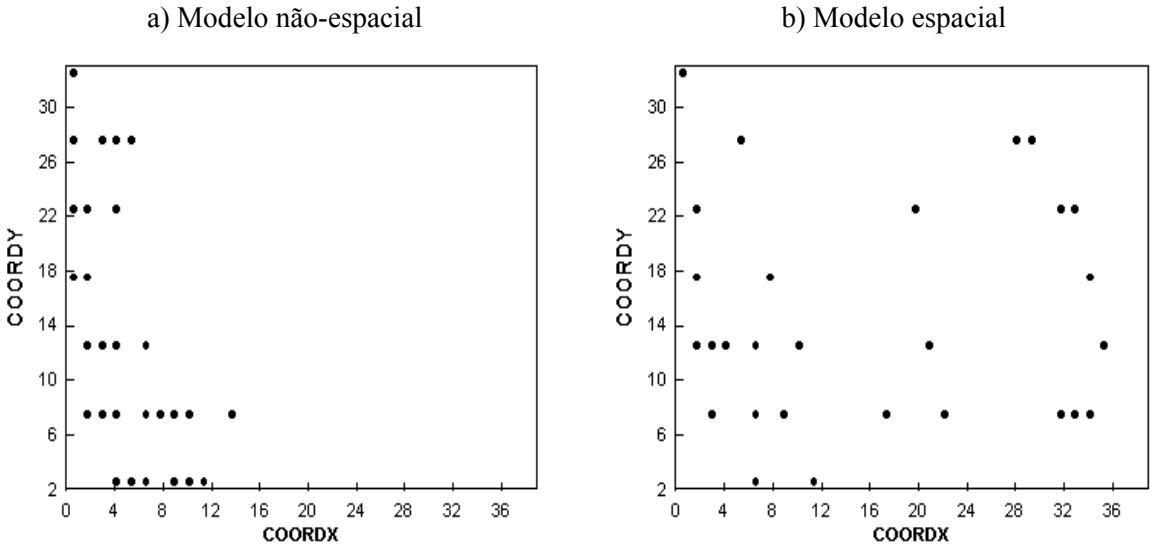

Figura 9. Localização das parcelas com as linhagens mais produtivas (25%) de um total de 110 avaliadas (genótipos não replicados), segundo dois modelos de análise estatística, (a) e (b), num ensaio de linhagens de soja, em blocos aumentados (Piracicaba, 1994/95).

5 CONSIDERAÇÕES FINAIS

As constatações aqui levantadas reforçam a preocupação acerca do problema da especificação de modelos de análise estatística, na área de melhoramento vegetal. Os três casos relatados, entre outros descritos na literatura (Stroup *et al.*, 1994; Reis & Miranda Filho, 2003), evidenciam que as dificuldades de se obter ganhos com a seleção, sobretudo quando se trabalha com espécies já bastante melhoradas (soja, milho, arroz, cana-de-acúcar etc), podem ser decorrentes do emprego de procedimentos estatísticos de menor poder de discriminação genotípica.

Considerando o pequeno número dos genótipos que efetivamente são liberados como cultivares comerciais, a adoção de abordagens analíticas menos restritivas e mais adequadas à natureza dos dados pode, portanto, determinar o sucesso ou o fracasso de um programa de melhoramento genético. Por fim, vale ressaltar que essa lacuna, claramente identificada na área de melhoramento de plantas, representa uma oportunidade de efetiva contribuição para os profissionais de estatística aplicada à experimentação agronômica.

6 REFERÊNCIAS

BESAG, J.; KEMPTON, R. Statistical analysis of field experiments using neighbourind plots. **Biometrics**, v. 42, p. 231 – 251, 1986.

BUENO FILHO, J.S. de S. Modelos mistos na predição de valores genéticos aditivos em testes de progênies florestais. Piracicaba, 1997. 118 p. Tese (Doutorado) - ESALQ/USP.

CULLIS, B.R.; LILL, W.J.; FISHER, J.A.; READ, B.J. A new procedure for the analysis of early generation variety trials. **Appl. Statist.**, v. 38, p. 361 – 375, 1989.

DUARTE, J. B. Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal. Piracicaba, 2000. 293 p. Tese (Doutorado) – ESALQ / USP. <disponível em: http://www.teses.usp.br>

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

DUARTE, J. B.; VENCOVSKY, R. Spatial statistical analysis and selection of genotypes in plant breeding. **Pesq. Agrop. Brasileira**, v. 40, n. 2, p. 107 – 114, 2005.

ES, H. M. van & ES, C. L. van. Spatial nature of randomization and its effect on the outcome of field experiments. **Agronomy J.**, v. 85, p. 420 – 428, 1993.

FEDERER, W.T. Augmented (or hoonuiaku) designs. **Hawaiian Planter's Records**, v. 55, p. 191 – 208, 1956.

FEDERER, W.T.; WOLFINGER, R.D. SAS code for recovering intereffect information in experiments with incomplete block and lattice rectangle designs. **Agronomy J.**, v. 90, p. 545 – 551, 1998.

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.

GRONDONA, M.O; CRESSIE, N. Using spatial considerations in the analysis of experiments. **Technometrics**, v. 33, p. 381 – 392, 1991.

HENDERSON C.R. **Applications of linear models in animal breeding**. Guelph: University of Guelph - Canada, 1984. 462 p.

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

PIEPHO, H.P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. **Theor. Appl. Genetics,** v. 89, p. 647 - 654, 1994.

REIS, A.J. S.; MIRANDA FILHO, J. B. Autocorrelação espacial na avaliação de compostos de milho para resistência à lagarta do cartucho (*Spodoptera frugicperda*). **Pesq. Agrop. Tropical**, v. 33, n. 2, p. 65 – 72, 2003.

SEARLE, S. R. **Linear models for unbalanced data**. New York: John Wiley & Sons, 1987. 536 p.

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

STROUP, W.W.; MULITZE, D.K. Nearest neighbor adjusted best linear unbiased prediction. **The American Statistician**, v. 45, n. 3, p. 194 – 200, 1991.

STROUP, W.W.; BAENZIGER, P.S.; MULITZE, D.K. Removing spatial variation from wheat yield trials: a comparison of methods. **Crop Sci.**, v. 34, p. 62 – 66, 1994.

WOLFINGER, R.D.; FEDERER, W.T.; CORDERO-BRANA, O. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. **Agronomy J.**, v. 89, p. 856 – 859, 1997.

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.