CAPÍTULO 5

APLICAÇÃO DE ANÁLISE ESTATÍSTICA ESPACIAL A EXPERIMENTOS GENÉTICOS EM BLOCOS AUMENTADOS

RESUMO

O presente trabalho propôs-se a ilustrar a aplicação de uma análise estatística espacial a experimentos genéticos delineados em blocos aumentados. Buscou-se, com isso, demonstrar os benefícios desta abordagem quando as observações experimentais não são espacialmente independentes. O material básico consistiu de um ensaio de competição de linhagens de soja, com cinco cultivares testemunhas (de efeitos fixos) e 110 novos genótipos (de efeitos aleatórios). O ajuste espacial foi feito pelo *modelo linear de campo aleatório* (*RFLM*), com função de autocovariância estimada a partir dos resíduos da análise sob erros independentes. Os resultados apontaram uma autocorrelação residual de magnitude e alcance significativos, o que garantiu à abordagem espacial uma melhoria considerável na discriminação dos tratamentos genéticos (aumento do poder dos testes estatísticos, redução nos erros padrão de estimativas e preditores, e alargamento na amplitude das predições genotípicas). Além disso, a análise espacial levou a um diferente ordenamento das linhagens em relação à análise não espacial e, finalmente, a uma seleção menos influenciada por efeitos da variação local. Tais diferenças podem ter implicações importantes no resultado final de programas de melhoramento como o aqui considerado.

***Palavras-chave**: delineamentos aumentados, análise espacial, informação intergenotípica, modelo misto, dados correlacionados, modelo linear de campo aleatório, autocovariância espacial.*

APPLICATION OF SPATIAL STATISTICAL ANALYSIS TO VARIETY TRIALS IN AUGMENTED BLOCK DESIGN

ABSTRACT

The aim of this work was to illustrate the application of a spatial statistical analysis to variety trials designed in augmented block and to demonstrate the benefits of this approach when the experimental observations are not spatially independent. The basic material of this study was an evaluation trial of soybean lines, with five check varieties (of fixed effect) and 110 new genotypes (of random effects). The spatial adjustment was made by *random field linear model* (*RFML*), with covariance function estimated from the residuals of the analyses under independent errors. Results showed a residual autocorrelation of significant magnitude and extension (*range*), which allowed a greater discrimination among genotypes when the spatial analysis was applied (increase of the power of statistical tests, reduction in the standard error of estimates and predictors, and increment of the amplitude of predictor values). Furthermore, the spatial analysis led to a different ranking of the genetic materials, in comparison with the nonspatial analysis, and a selection less influenced by effects of local variation was obtained. Such differences may have important consequences on the final outcome of breeding programs as this one here considered.

Key words*: augmented designs, spatial analysis, intergenotypic information, mixed model, correlated data, random field linear model, spatial autocovariance.*

1. INTRODUÇÃO

Uma característica marcante das fases preliminares dos programas de seleção, no melhoramento de plantas, é a pequena quantidade de material de propagação para cada novo genótipo a ser avaliado. Isto limita o uso de repetições para estes tratamentos genéticos, os quais são, com freqüência, avaliados numa só parcela experimental (sem repetições). Para lidar com esse tipo de limitação Federer (1956; 1958) 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*) e, portanto, na suposição de independência entre observações.

Por outro lado, a pouca disponibilidade de material (sementes, tubérculos, etc.) 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 isto, infelizmente, aumenta a chance de violação da independência entre observações assumida pelo método *OLS*, haja vista a provável maior similaridade ou correlação entre observações de parcelas vizinhas (Stroup *et al*., 1994). Este fenômeno, denominado *autocorrelação espacial*, pode comprometer seriamente a comparação de tratamentos. Em experimentos simulados, Es & Es (1993) demonstraram que, sob esta correlação, os testes estatísticos associados a contrastes de tratamentos cujas parcelas estiveram separadas por pequenas distâncias têm maior probabilidade de *erro tipo II* (tratamentos diferentes apontados como estatisticamente iguais). Enquanto os contrastes de tratamentos cujas parcelas estiveram separadas por distâncias maiores foram testados com maior probabilidade de *erro tipo I* (tratamentos iguais apontados como estatisticamente diferentes).

Além disso, nos ensaios em blocos, a autocorrelação espacial compromete as suposições de homogeneidade intrablocos e de ausência da interação 'tratamentos *x* blocos', fundamentais para a eficiência destes delineamentos (Gusmão, 1986; Warren & Mendez, 1982). Neste sentido, o uso de blocos grandes (com mais de oito ou doze parcelas), freqüente nos experimentos genéticos com elevado número de tratamentos, também contribui para a violação destas suposições e a conseqüente ineficiência dos delineamentos (Stroup *et al*., 1994). Warren & Mendez (1982) acrescentam que os blocos também falham no controle de gradientes ou tendências da variabilidade espacial quando não são adequadamente orientados e/ou possuem formato desfavorável. A maioria desses problemas é comum aos ensaios de blocos aumentados na área de melhoramento de plantas, podendo prejudicar sensivelmente a qualidade das estimativas genotípicas e inflacionar a variância

do erro experimental. Por conseguinte, a detecção das diferenças reais entre os numerosos genótipos torna-se dificultada, sobretudo quando seus efeitos individuais são pequenos (populações com baixa variabilidade genotípica).

Deve-se ressaltar que, embora a análise de variância tradicional associada aos delineamentos clássicos ignore esse tipo de correlação, tais planejamentos pressupõem casualização, à qual caberia a tarefa de neutralizar esses efeitos prejudiciais (Brownie *et al*., 1993). Mas, segundo Stroup *et al*. (1994) tem-se constatado que as análises convencionais, nos delineamentos em blocos, freqüentemente não neutralizam de forma adequada a variabilidade espacial. Felizmente, avanços recentes em estatística para dados espacialmente distribuídos têm fornecido métodos alternativos mais eficientes nesse tipo de situação (Pithuncharurnlap *et al*., 1993). Por isso, Kempton *et al.* (1994) advogam, para a melhoria da precisão dos ensaios varietais, uma maior utilização dos métodos que consideram algum ajustamento para a heterogeneidade espacial.

A literatura é rica em trabalhos que propõem e discutem métodos estatísticos para a análise de dados experimentais espacialmente correlacionados (Papadakis,1937; Williams, 1952; Bartlett, 1978; Kirk *et al*., 1980; Kempton & Howes, 1981; Wilkinson *et al.*, 1983; Green *et al*., 1985; Besag & Kempton, 1986; Williams, 1986; Gleeson & Cullis, 1987; Tamura *et al*., 1988; Cullis *et al.*, 1989; Martin, 1990; Cullis & Gleeson, 1991; Stroup & Mulitze, 1991; Grondona & Cressie, 1991; Zimmerman & Harville, 1991; Eisenberg *et al*., 1996; Gleeson, 1997; Cullis *et al.*, 1998; Pearce, 1998; Federer, 1998). Alguns métodos levam em conta a correlação espacial em uma única direção (ajustamento unidimensional), geralmente no sentido da largura das parcelas (ex: Gleeson & Cullis, 1987; Cullis *et al.*, 1989). Outros consideram-na em duas direções (ajustamento bidimensional), aplicando-se a ensaios com parcelas de formato aproximadamente quadrado (ex: Martin, 1990; Cullis & Gleeson, 1991; Cullis *et al.*, 1998). Quase todas as propostas foram desenvolvidas para experimentos com repetições, exceto as de Cullis *et al.* (1989) e Cullis *et al.* (1998) que adaptaram essas técnicas para os ensaios preliminares de melhoramento (com uma só parcela por genótipo-teste e variedades testemunhas repetidas, as quais podem ser sistematicamente alocadas). Em síntese, a maioria desses métodos baseia-se num modelo de "tendência + erro" e emprega, pelo menos implicitamente, alguma forma de diferenciação dos dados (tomada de diferenças sucessivas baseadas no princípio de vizinhança) para remover a tendência assumida (Pithuncharurnlap *et al*., 1993). As diferenças entre eles estão, sobretudo, nas suas suposições e nos métodos de estimação associados ao efeito de tendência (Loo-Dinkins, 1992).

Uma abordagem um tanto distinta é a de Zimmerman & Harville (1991). Em sua proposta, os autores modelam diretamente o efeito de parcela (tendência + erro), de forma que as observações são consideradas coletivamente como uma realização parcial de um *campo aleatório* (Martínez, 1994). Neste caso, os efeitos de parcela são assumidos distribuírem-se de acordo com algum modelo de correlação espacial que descreve as tendências locais, análogos aos modelos de predição usados em geoestatística. Assim, como na maioria dos outros métodos, o modelo busca uma estimativa da função geral de covariância, a qual participa diretamente dos processos de estimação, predição, etc., via *quadrados mínimos generalizados* (*GLS*). Enfim, trata-se essencialmente de um modelo linear misto com erros espacialmente correlacionados (Stroup *et al*., 1994). Dada a sua concepção, os seus propositores chamaram-no de "modelo linear de campo aleatório" (*random field linear model – RFLM*). Sua vantagem comparativa está no fato de aplicar-se a ensaios com dependência espacial em todas ou quaisquer direções, em associação com os diversos esquemas de blocagem (ou nenhum), podendo ainda acomodar diferentes tamanhos e formas de parcelas. Ademais, a abordagem lida naturalmente com parcelas limítrofes (nas extremidades da área experimental), tornando-a livre de ambigüidades.

Diante do que foi exposto, entende-se que é realmente necessário difundir a abordagem estatística da análise espacial de experimentos no contexto dos programas de melhoramento. Assim, o objetivo deste trabalho é ilustrar o enfoque geoestatístico aplicado a este tipo de análise (*RFLM*), adaptando-o ao delineamento de blocos aumentados. Neste sentido, tal enfoque não tem o compromisso de representar a melhor abordagem espacial para o conjunto de dados analisados. Mas, sobretudo, o de demonstrar os benefícios potenciais de um tratamento estatístico menos restritivo em comparação ao tradicional enfoque de observações espacialmente independentes (não necessariamente *OLS*). Os procedimentos estatístico-computacionais são descritos basicamente em linguagem *SAS®* (*Statistical Analysis System*), tendo em vista o propósito de facilitar a sua implementação em programas de melhoramento.

2. MATERIAL E MÉTODOS

2.1. Material

Os dados experimentais utilizados neste estudo provêm de um ensaio de competição de linhagens ($F_{6:3}$) de soja, conduzido no local Areão, município de Piracicaba-SP, em 1994/95. O ensaio faz parte do programa de seleção recorrente visando o aumento da produtividade de grãos na soja, desenvolvido no Departamento de Genética da ESALQ/USP (Setor de Genética Aplicada às Espécies Autógamas). Mais especificamente, o experimento avaliou um conjunto de progênies

resultantes de cruzamentos biparentais entre cultivares do grupo de maturação semi-precoce. O experimento foi delineado em blocos aumentados, com *t=5* testemunhas (Bossier, Davis-1, IAC-12, IAS-5 e Viçoja) e, inicialmente, *p=180* tratamentos adicionais (progênies), distribuídos em *b=4* blocos de aproximadamente 50 parcelas. A unidade experimental correspondeu a duas fileiras de plantas, espaçadas 0,6m entre si e com 5m de comprimento. Embora vários caracteres tenham sido avaliados, apenas os dados de produtividade de grãos (kg/ha) foram aqui considerados. Dada a perda de parcelas para este caráter e a eliminação daquelas de estande muito baixo (observações discrepantes), com vistas à distribuição normal dos dados, restaram, finalmente, *p=110* progênies, totalizando *n=127* observações (Tabela 5.1, em Anexos).

O ensaio foi escolhido por apresentar indícios de autocorrelação residual e por mostrar um elevado coeficiente de variação (CV=53%) quando submetido à análise intrablocos (*OLS*). Para a implementação da análise estatística espacial fez-se necessário ainda determinar as distâncias, em metros, correspondentes às coordenadas geográficas do centro de cada parcela na grade de campo do experimento. Denotou-se, então, **`COORDX`**, a coordenada no sentido da largura das parcelas, e **`COORDY`**, a coordenada no sentido do comprimento das parcelas (Tabela 5.1).

2.2. Procedimentos estatístico-computacionais

2.2.1. Os modelos de análise estatística

Os dados experimentais foram submetidos às análises estatísticas relacionadas a dois modelos matemáticos: *i)* modelo assumindo observações espacialmente independentes; e *ii)* modelo admitindo correlação espacial entre observações. Dado o elevado número de progênies (tratamentos adicionais) e uma suposta população de referência, os efeitos destes tratamentos foram admitidos como aleatórios. Logo, a análise de erros independentes aqui implementada não corresponde a de modelo fixo (*OLS*). Ambos são modelos mistos com tratamentos adicionais aleatórios, assumidos independentes e oriundos de uma única população, a qual é supostamente de natureza fixa tal como as testemunhas. A diferença nos dois modelos restringe-se, portanto, à suposição associado ao erro experimental.

Esse tipo de modelagem foi descrito para a situação de observações espacialmente independentes como análise com recuperação de *informação intergenotípica* (Wolfinger *et al*., 1997; Federer, 1998). Sua peculiaridade, no caso dos delineamentos aumentados, é que o modelo matemático deve acomodar efeitos de tratamentos (genótipos) de duas naturezas, fixos para as testemunhas e aleatórios para as progênies (dentro da *(t+1)–ésima* população fixa). Logo, para a

primeira alternativa de análise (*i*), as observações podem ser individualmente caracterizadas pelo seguinte modelo (adaptado de Scott & Milliken, 1993):

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

em que:

Y_{ijk}: é a resposta observada 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 $N(0, \sigma^2_g)$ independente, se o genótipo for um tratamento adicional *(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 $N(0, \sigma^2_e)$.

Já no modelo de análise que admite autocorrelação espacial (*ii*), o termo e_{ijk} é assumido como: $e_{ijk} \sim N[0, C(h)]$; sendo $C(h)$ a (co)variância entre dois erros de parcelas separadas por uma distância *h*, com $h \geq 0$ e os erros denotados por $e_{(s)}$ e $e_{(s+h)}$ (*s* indica a posição espacial da *ijk-ésima* parcela). Na abordagem de campo aleatório (*RFLM*), $C(h)$ é definida como (Littell *et al*., 1996):

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

Logo, a covariância dos erros é assumida ser uma função da distância que separa as correspondentes parcelas ($f(h)$). Esta função, entretanto, não é estabelecida previamente, mas é estimada do "ensaio de uniformidade" sugerido pelos resíduos do ajuste do modelo com erros independentes ($\hat{e}_{ijk}$).

De uma forma genérica, expressando-se as observações por um vetor $\mathbf{y}$, ambos os modelos podem ser matricialmente representados pelo *modelo linear misto geral* (Henderson, 1984):

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

Neste caso: os efeitos fixos são reunidos no vetor paramétrico $\boldsymbol{\beta}$; os efeitos aleatórios, no vetor paramétrico $\boldsymbol{\gamma}$, exceto os erros que compõem o vetor $\boldsymbol{\varepsilon}$; e, $\mathbf{X}$ e $\mathbf{Z}$ representam as matrizes de incidência dos efeitos contidos em $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente. Os efeitos genotípicos aleatórios ($\boldsymbol{\gamma}$) são aqui assumidos (sem perda de generalidade do método) terem distribuição normal com média nula ($\boldsymbol{\phi}$) e matriz de (co)variâncias $\mathbf{G}=\mathbf{I}\sigma^2_g$ (onde $\mathbf{I}$ é uma matriz identidade). Os erros experimentais

têm, supostamente, distribuição normal com média nula e matriz genérica de (co)variâncias **R**. Assim, no modelo que pressupõe independência espacial entre observações tem-se: $\mathbf{R}=\mathbf{I}\sigma_e^2 \Rightarrow \mathbf{V}=\mathbf{ZZ}'\sigma_g^2+\mathbf{I}\sigma_e^2$ (num modelo fixo isto implicaria na independência completa das observações: $\mathbf{V}=\mathbf{I}\sigma_e^2$). Já no modelo que acomoda observações espacialmente correlacionadas, a matriz $\mathbf{R}=\mathbf{\Sigma}$ não é mais diagonal e sua estrutura dependerá do alcance da referida autocorrelação, bem como da função geral de covariância utilizada para descrever a dependência espacial (item 2.2.3).

2.2.2. Ajustando o modelo de observações espacialmente independentes

O primeiro passo das análises estatísticas aqui implementadas é o ajustamento do modelo que pressupõe independência espacial entre as observações. Para fins de aplicação, as instruções estatístico-computacionais são apresentadas em linguagem *SAS*. O arquivo de dados deve, então, ser construído similarmente à Tabela 5.1, acrescentando-se, no momento de sua leitura (criação do *dataset SAS*), duas novas variáveis, '**C**' e '**NEW**', conforme as instruções a seguir:

```
data TAB_51;
  infile 'C:\DADOS\TAB_51.PRN';
  input PARCELA COORDX COORDY BLOCO GENOT$ TIPO$ PG;
  C=GENOT; if TIPO='New' then C=0;
  if TIPO='New' then NEW=1; else NEW=0;
run;
```

O ajuste do modelo pode, então, ser obtido utilizando-se o procedimento do *SAS* para análise de modelos lineares mistos (*PROC MIXED*), e um conjunto de instruções como:

```
proc mixed data=TAB_51;
  class BLOCO C GENOT;
  model PG=BLOCO C / ddfm=satterth p;
  random GENOT*NEW /solution cl;
  id PARCELA COORDX COORDY BLOCO GENOT TIPO;
  contrast 'TESTEMUNHAS'  C  0 1 -1 0 0 0,
                          C  0 0 1 -1 0 0,
                          C  0 0 0 1 -1 0,
                          C  0 0 0 0 1 -1;
  contrast 'TESTs vs PROGs' C -5 1 1 1 1 1;
  contrast 'PROG1 vs PROG2' | GENOT*NEW 1 -1;
  contrast 'PROG2 vs PROG3' | GENOT*NEW 0 1 -1;
  lsmeans C / cl;
  make 'solutionr' out=EBLUPs;
  make 'predicted' out=OBS_PRED;
  make 'fitting' out=R_IND (rename=(value=VAL_RI));
run;
```

Numa breve descrição, os comandos utilizados neste programa têm as seguintes funções (SAS Institute, 1997): '**proc mixed**' invoca o procedimento, o qual processa o *dataset* '**TAB_51**', utilizando o método de máxima verossimilhança restrita – *REML* (método *default*) para a estimação dos componentes de variância (σ_g^2 e σ_e^2); '**class**' lista as variáveis de classificação no modelo; '**model**' informa a variável resposta e os efeitos fixos relacionados, acrescidos, aqui, das opções para aproximar os graus de liberdade das estatísticas *F* e *t* por Satterthwaite (**ddfm=satterth**) e para listar os valores preditos e residuais das observações (**p**); '**random**' lista os efeitos aleatórios, neste caso, adaptado para excluir os efeitos das testemunhas (Wolfinger *et al*., 1997), e com as opções para listar os correspondentes *EBLUP*'s (*empirical best linear unbiased predictors*) e seus intervalos de confiança[1] (**solution** e **cl**, respectivamente); '**id**' mantém as referidas variáveis na listagem de valores preditos; '**contrast** ...' fornece o teste *F* para contrastes específicos de efeitos fixos ou aleatórios (as progênies são ordenadas numérica e alfabeticamente); '**lsmeans**' produz médias ajustadas de efeitos fixos, neste caso, as médias associadas à variável '**C**' ($\hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_t$, sendo $\hat{\mu}_0$ a estimativa da média da população de progênies); e, '**make** ...' cria *dataset*'s específicos de interesse em processamentos posteriores (**R_IND** armazenará estatísticas relacionadas ao ajuste do modelo; **EBLUPs**, os preditores dos efeitos aleatórios; e **OBS_PRED**, os valores preditos e residuais das observações).

O primeiro arquivo (**R_IND**) conterá resultados que são úteis para a comparação estatística dos dois modelos de análise. O segundo possibilita o ordenamento das progênies para fins de seleção, o que pode ser obtido executando-se a seguinte instrução:

```
proc sort data=EBLUPs; by descending _est_;
proc print; run;
```
[2]

O último arquivo (**OBS_PRED**) terá os resíduos do ajuste do modelo de observações espacialmente independentes, os quais são úteis para avaliar e estimar a estrutura de correlação espacial presente nos dados. Os resíduos ($\hat{e}_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$) podem ser expressos vetorialmente por: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$, com $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^0 + \mathbf{Z}\tilde{\boldsymbol{\gamma}}$; sendo: $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ e $\tilde{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$ (SAS Institute, 1997; Henderson, 1984).

---

[1]/ O nível de confiança *default* é 95% (modificações através da opção '**alpha=** ').

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

2.2.3. Diagnosticando e estimando a covariância espacial

A autocorrelação residual foi avaliada por dois tipos de instrumentos: o teste estatístico de Durbin-Watson (*d)* e algumas representações gráficas dos resíduos. A estatística *d*, que permite testar a hipótese de ausência de autocorrelação (H_0: $\rho=0$), é definida como (Hoffmann & Vieira, 1998; SAS Institute, 1993a; Gujarati, 1992):

$$d = \frac{\sum_{l=2}^{n}(\hat{e}_l - \hat{e}_{l-1})^2}{\sum_{l=1}^{n}\hat{e}_l^2}$$

sendo: $l=1, 2,..., n$, a ordem de posicionamento da parcela associada ao resíduo $\hat{e}_l$ ($\hat{e}_l$ e $\hat{e}_{l-1}$ indicam resíduos cujas parcelas têm vizinhança de primeira ordem, isto é, são adjacentes) .

A relação entre *d* e ρ é aproximadamente: $d=2(1-\rho)$. Logo, se não existir autocorrelação o valor esperado de *d* é *2,0*; valores significativamente inferiores a *2,0* indicam autocorrelação positiva; e, valores significativamente superiores a *2,0* indicam autocorrelação negativa. Embora o teste tenha sido delineado, em princípio, para autocorrelação de primeira ordem, o *SAS* permite, através do *PROC AUTOREG* (procedimento para o ajuste de modelos auto-regressivos), obter estatísticas de Durbin-Watson generalizadas até uma dada ordem especificada. As instruções aqui utilizadas para testar autocorrelações residuais de primeira à décima ordem (**dw=10**) foram:

```
proc autoreg data=OBS_PRED;
  model _resid_= / dw=10 dwprob;
run;
```

O primeiro tipo de gráfico de resíduos compreendeu a dispersão dos valores $\hat{e}_{ijk}$ (ou $\hat{e}_l$) em função do número das respectivas parcelas no campo experimental ($l=1,2,...,n$). As parcelas foram identificadas seqüencialmente no chamado *sentido do caminhamento* (ziguezague). Inspecionou-se também o comportamento dos resíduos em função das coordenadas X e Y. Nestes gráficos, configurações em que os resíduos apresentam distribuição aleatória sugerem ausência de autocorrelação, enquanto aquelas em que resíduos seqüenciais (vizinhos) tendem a se agrupar caracterizam a presença de autocorrelação serial (Draper & Smith, 1981; Parente, 1984).

Outro tipo de gráfico que também serve para diagnosticar a presença de autocorrelação é o *semivariograma* ou simplesmente *variograma*. Aqui, como na geoestatística, este gráfico foi utilizado, sobretudo, para a estimação da estrutura de variabilidade espacial (nos resíduos) como função da distância entre observações (Littell *et al*., 1996). Nesta representação, valores estimados de semivariância ($\hat{S}(h)$) são plotados contra suas respectivas distâncias (h), resultando no chamado

variograma amostral. A semivariância é definida como: $S(h)=\frac{1}{2}Var[\,e_{(s+h)}-e_{(s)}\,]$ (Stroup *et al*., 1994); e seu correspondente estimador, neste estudo, foi:

$\hat{S}(h)=\frac{1}{2N(h)}\sum_{N(h)}[\hat{e}_{(s+h)}-\hat{e}_{(s)}]^2$; sendo $N(h)$ o número de diferenças tomadas à distância h.

Em síntese, um variograma amostral é obtido pelos seguintes passos: *1°)* fixa-se uma distância h ("*lag*"); *2°)* formam-se todos os pares de observações separadas pela distância h; *3°)* aplica-se a expressão do estimador, obtendo-se a semivariância associada à distância h; *4°)* toma-se outra distância (*lag*) e repete-se os passos de *1* a *3*, o que é feito até uma distância máxima com razoável número de pares; *5°)* num gráfico com distâncias no eixo das abscissas e semivariâncias no eixo das ordenadas plotam-se, então, os pontos amostrados $(h,\hat{S}(h))$ (Ribeiro Júnior, 1995; Vieira, 2000). Neste gráfico, valores de $\hat{S}(h)$ distribuindo-se aleatoriamente em função de h refletem independência das observações (resíduos). Mas, se os valores $\hat{S}(h)$ tendem a crescer com o aumento de h até uma determinada distância (*alcance*), a partir da qual a semivariância se estabiliza (*patamar*), tem-se a configuração típica de dependência espacial entre observações (menor variabilidade associada às distâncias menores). O *alcance* da correlação espacial (a) é a distância média de influência de uma observação (parcela), assumido, aqui, uniforme em todas as direções (*isotropia*). Já o *patamar* (σ^2), às vezes referido como "*sill*", corresponde à variância intrínseca da variável em estudo ($Var[e_{(s)}]=Cov[e_{(s)},e_{(s)}]$); ou ainda, à covariância entre resíduos de parcelas separadas por uma distância igual ou superior ao alcance ($Cov[e_{(s)},e_{(s+h)}]$, com $h\geq a$).

A vantagem de se avaliar a dependência espacial através do variograma é que, sob estacionariedade[3], este tem uma relação direta e simples com a função de autocovariância ($C(h)$): $S(h)=\sigma^2-C(h)$; sendo $\sigma^2=C(h=0)$ (Stroup *et al*., 1994; Es & Es, 1993; Grondona & Cressie, 1991; Pannatier, 1996; Vieira, 2000). Séries não estacionárias, em geral, tornam-se estacionárias removendo-se alguma tendência nos dados que seja função de suas respectivas localizações, o que representa, inclusive, uma possibilidade de "controle local" *a posteriori* propiciada pela abordagem espacial. No presente trabalho não se buscou esse tipo de remoção. Primeiramente porque os erros

[3]/ Uma função aleatória $Z(s)$ possui estacionariedade se a sua lei espacial é invariante à translação, isto é: $\{Z(s_1), Z(s_2), ..., Z(s_k)\}$ e $\{Z(s_1+h), Z(s_2+h), ..., Z(s_k+h)\}$ têm a mesma lei de distribuição para qualquer vetor h de separação. Em geoestatística linear é suficiente uma estacionariedade de segunda ordem ($E\{Z(s)\}=m$, $\forall\ s$; e, $C(h)=E\{Z(s).Z(s+h)\}-m^2$, $\forall\ s$), condição que normalmente é atendida desde que exista um variograma típico (com uma parte crescente e um patamar) e este seja estacionário – *hipótese intrínseca* (Pannatier, 1996; Vieira, 2000).

sugeriram estacionariedade (variando em torno de uma média fixa – Gleeson, 1997), e também porque isto prejudicaria a legítima comparação entre as duas estruturas de erros ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$ e $\mathbf{R}=\mathbf{\Sigma}$), em modelos com idênticas fontes de variação (mesmas matrizes $\mathbf{X}$ e $\mathbf{Z}$). Portanto, como em Littell *et al*. (1996), efeitos espaciais aleatórios e de possíveis tendências foram modelados conjuntamente.

Diante disso, ajustando-se uma função contínua ao variograma amostral (descontínuo) obtém-se, pela relação $C(h)=\sigma^2-S(h)$, a correspondente função que descreve a covariância espacial. Naturalmente, o uso de uma função única para modelar a dependência espacial em todas as direções está fundamentado na suposição simplificada de que o processo é *isotrópico*, ou seja, a lei espacial de formação dos resíduos é supostamente a mesma em qualquer direção, sendo, por isso, estimada por um só variograma (*omnidirecional*)[4]. Isto não necessariamente ocorre, mas, representa uma aproximação satisfatória da realidade em várias situações práticas (Zimmerman & Harville, 1991). As funções de variograma mais comumente utilizadas são os modelos denominados: *esférico*, *exponencial* e *gaussiano*, aos quais correspondem, para campos aleatórios isotrópicos, as seguintes funções de autocovariância (SAS Institute, 1997; Grondona & Cressie, 1991):

Modelo esférico:

$$C(h)=\begin{cases}\sigma^2\left[1-\frac{3}{2}\left(\frac{h}{a}\right)+\frac{1}{2}\left(\frac{h}{a}\right)^3\right], \text{ se } & h<a\\ 0, \text{ se } & h\geq a\end{cases}$$

Modelo exponencial:

$C(h)=\sigma^2\ exp\left(\frac{-h}{a_e}\right)$, em que: $a_e=\frac{1}{3}a$ (Webster, 1985).

Modelo gaussiano:

$$C(h)=\sigma^2\ exp\left(\frac{-h^2}{a^2}\right)$$

Dada a ampla aplicação do variograma em geoestatística, são disponíveis *softwares* que facilitam sobremaneira o seu ajustamento (ex: *VARIOWIN*; *Geo-EAS*). Um consenso no ajuste desses modelos é que eles devem, preferencialmente, ajustar-se melhor aos pontos correspondentes às menores distâncias. Isso porque, na descrição da similaridade espacial entre observações, uma estimativa de semivariância é cada vez menos importante à medida que aumenta a sua respectiva distância. Evidências empíricas também demonstram que apenas as correlações amostrais para pequenas distâncias entre parcelas (*lags*) necessitam ser levadas em conta nos modelos de análise espacial (Gleeson, 1997; Gleeson & Cullis, 1987). Por isso, na prática, os métodos de quadrados mínimos (não lineares) geralmente são preteridos ao chamado "ajuste a sentimento". Por este processo, a busca da função que melhor se ajusta aos pontos do variograma amostral é, via de regra,

---

[4]/ Processos espaciais *anisotrópicos* exigem a modelagem de dois ou mais variogramas *direcionais*.

obtida por meio de um *software* que permite avaliar diferentes atribuições aos valores paramétricos σ^2 e a. Neste trabalho, utilizou-se o aplicativo *VARIOWIN 2.2* (Pannatier, 1996), o qual permite, através de barras de rolagem, modificar suavemente o valor de cada parâmetro, acompanhando simultaneamente o ajustamento gráfico. Embora este procedimento mostre uma certa subjetividade, pequenas variações nos valores destes parâmetros normalmente não modificam as conclusões finais da análise do experimento. Ademais, Zimmerman & Harville (1991) asseguram que as estimativas de contrastes de tratamentos são relativamente robustas a qualquer escolha entre as funções de covariância aqui apresentadas, todas contínuas, não negativas e monotônicas decrescentes.

Littell *et al*. (1996) apresentam um programa *SAS* que permite, através do *PROC MIXED*, buscar as estimativas *REML* da estrutura de covariância espacial. O procedimento, naturalmente, também usa os resíduos do ajuste do modelo com observações espacialmente independentes (*dataset* **OBS_PRED**). O conjunto de instruções adaptadas para o caso presente é:

```
proc mixed data=OBS_PRED;
  model _resid_= ;
  repeated / sub=intercept type=sp(EXP) (COORDX COORDY);
  parms (0 to 30 by .5) (100000 to 200000 by 5000);
run;
```

Essas estimativas, entretanto, não são obtidas priorizando-se as correlações amostrais relacionadas às pequenas distâncias, o que pode comprometer a eficiência da análise espacial para grandes conjuntos de dados (Littell *et al*., 1996). Ademais, neste programa, é necessário especificar o tipo da função de covariância espacial ('**type=sp(EXP)**' exemplifica a escolha do modelo exponencial) e os limites entre os quais o *PROC MIXED* deve buscar as estimativas dos parâmetros a e σ^2, respectivamente ('**parms (0 to 30 by .5)** ...' exemplifica uma busca de a entre *0* e *30*m, com incremento de *0,5*m; e assim por diante). Estas informações, sem dúvida, são melhor aproximadas quando se ajusta também o variograma. Enfim, um trabalho paralelo com os dois procedimentos pode reduzir os problemas inerentes a ambas as formas de ajustamento. No presente trabalho, embora isso tenha sido feito, priorizaram-se as estimativas obtidas a partir do modelo de variograma, conforme recomendam os autores anteriormente referidos.

2.2.4. Ajustando o modelo de análise espacial

Uma vez definido o tipo da função de autocovariância e os correspondentes valores de σ^2 e a, que descrevem a estruturação dos resíduos (*etapa de caracterização*), a fase seguinte é a de *ajustamento* do modelo que admite dependência espacial entre observações. Essa etapa envolve a obtenção de estimativas, preditores e testes estatísticos relacionados aos tratamentos, que sejam

livres dos efeitos da correlação espacial estimada. As instruções *SAS* utilizadas para este ajustamento foram (adaptadas de Littell *et al*., 1996):

```
proc mixed data=TAB_51 noprofile;
  class BLOCO C GENOT;
  model PG= BLOCO C / ddfm=satterth;
  random GENOT*NEW / solution cl;
  repeated / sub=intercept type=sp(EXP) (COORDX COORDY);
  parms (17242.02) (126450) (6.8) / noiter;
  contrast 'TESTEMUNHAS'  C  0 1 -1 0 0 0,
                          C  0 0 1 -1 0 0,
                          C  0 0 0 1 -1 0,
                          C  0 0 0 0 1 -1;
  contrast 'TESTs vs PROGs' C -5 1 1 1 1 1;
  contrast 'PROG1 vs PROG2' | GENOT*NEW 1 -1;
  contrast 'PROG2 vs PROG3' | GENOT*NEW 0 1 -1;
  lsmeans C / cl;
  make 'solutionr' out=EBLUPs;
  make 'fitting' out=R_ESP (rename=(value=VAL_RE));
run;
```

Neste programa, além do que já foi descrito anteriormente (item 2.2.2), merece destaque o comando '**repeated**' que especifica a estrutura da matriz **R** (a estrutura de **G** é definida em '**random**'[5]). Sua opção '**sub=intercept**' trata todas as observações como potencialmente correlacionadas; assim, se o experimento fosse conduzido em vários locais seria conveniente, por exemplo, fazer '**sub=LOCAL**' (Littell *et al*., 1996). A opção '**type=** ' permite escolher entre diversas estruturas de covariância disponíveis no *PROC MIXED*. Aqui, a escolha incidiu sobre uma estrutura espacial (**sp**), em duas dimensões (**COORDX** e **COORDY**), cuja função de autocovariância é o modelo exponencial (**EXP**). O comando '**parms**' especifica, então, os valores dos parâmetros de covariância associados aos efeitos listados em '**random**' e '**repeated**', respectivamente (neste caso a ordem é: σ_g^2, σ^2 e a_e). E, sua opção '**noiter**' associada a '**noprofile**' de '**proc mixed**' impedem qualquer iteração a partir destes valores. Por fim, os demais comandos possibilitam a obtenção de estimativas, preditores e testes estatísticos úteis na comparação dos tratamentos, bem como na avaliação da eficiência relativa dos dois modelos de análise.

---

[5]/ O tipo de estrutura da matriz **G** não é aqui informado pois esta tem, neste caso, a estrutura *default* $\mathbf{I}\sigma_g^2$. Da mesma forma, na análise sob observações espacialmente independentes ($\mathbf{R}=\mathbf{I}\sigma_e^2$), dispensou-se o uso do comando '**repeated**'.

Embora uma nova estimativa de σ_g^2 pudesse ser buscada (utilizando-se, por exemplo, '**`parms (10000 to 30000 by 500)` ...**' e suprimindo-se '**`noiter`**'), optou-se por manter o valor obtido do modelo de erros independentes (**`17242.02`**), para que se pudesse avaliar somente a influência das diferentes estruturas de $\mathbf{R}$ ($\mathbf{I}\sigma_e^2$ e $\mathbf{\Sigma}$). Logicamente, os parâmetros da estrutura de covariância espacial também poderiam ser assim estimados, o que implicaria na execução de uma só análise (estimação e predição simultâneas). Contudo, isto introduziria todos os inconvenientes já reportados, além de maiores requisitos computacionais.

2.2.5. Comparando os modelos de análise espacial e não espacial

A abordagem de modelos mistos baseada em verossimilhança fornece diversas medidas estatísticas que permitem comparar, quanto à adequação, modelos com diferentes estruturas de covariância. As mais difundidas são: os critérios de informação de Akaike (*AIC*) e de Schwarz (*BIC*), e o teste da razão de verossimilhança (*LRT*). Os dois primeiros são baseados no valor que maximiza o logaritmo da verossimilhança restrita ($l_{REML}(\mathbf{G},\mathbf{R})$), descontado de uma função do número de parâmetros de covariância (q). Dessa forma, o modelo com os maiores valores de *AIC* e *BIC* deve ser o preferido (detalhes teóricos em SAS Institute, 1997, p. 650). Acrescenta-se que ambos os critérios já estão implementados no *SAS* e fazem parte da saída padrão do *PROC MIXED* (tabela '*model fitting information*').

A outra opção (*LRT*) permite testar estatisticamente a diferença de adequação dos dois modelos (Littell *et al*., 1996): $\chi^2_{(f)} = [-2l_{REML}(\mathbf{G},\mathbf{R})_{\mathbf{I}}] - [-2l_{REML}(\mathbf{G},\mathbf{R})_{\mathbf{\Sigma}}]$, (os índices $_{\mathbf{I}}$ e $_{\mathbf{\Sigma}}$ remetem aos modelos com $\mathbf{R}=\mathbf{I}\sigma_e^2$ e $\mathbf{R}=\mathbf{\Sigma}$, respectivamente). Os valores $[-2l_{REML}(\mathbf{G},\mathbf{R})]$ são listados nas respectivas tabelas de informação das duas análises. Sob normalidade dos dados, o teste tem distribuição de qui-quadrado, neste caso, com apenas *um* grau de liberdade ($f=q_{\mathbf{\Sigma}}-q_{\mathbf{I}}=3-2$). Um programa *SAS* para resumir as informações do ajustamento dos modelos, bem como executar o referido teste é listado no Apêndice 5.1.

Outros critérios foram ainda utilizados na comparação dos dois modelos de análise. Um deles é a verificação da capacidade dos testes estatísticos captarem diferenças significativas, particularmente entre efeitos de tratamentos. Outro é a identificação de mudanças substanciais nas estimativas e predições relacionadas aos tratamentos genéticos, bem como no ordenamento dos genótipos com vistas à seleção. Neste sentido, avaliou-se também a influência dos dois modelos na precisão destas estatísticas.

3. RESULTADOS E DISCUSSÃO

3.1. Caracterização da covariância espacial

As Figuras de 5.1 a 5.3 ilustram o comportamento dos resíduos $\hat{e}_{ijk}$, decorrentes do ajuste do modelo que pressupõe observações espacialmente independentes. Na primeira delas (Figura 5.1), observa-se um nítido agrupamento entre resíduos vizinhos e uma certa periodicidade na sua distribuição ao longo do campo experimental. Esta configuração, por certo, resulta das autocorrelações positivas de primeira à sexta ordem detectadas na série de resíduos (Tabela 5.2), o que, conseqüentemente, comprova a dependência espacial entre observações. Em situações desse tipo, a adoção dos restritivos modelos tradicionais (assumindo $\mathbf{R}=\mathbf{I}\sigma_e^2$) implica numa redução da eficiência da análise estatística enquanto instrumento de tomada de decisões.

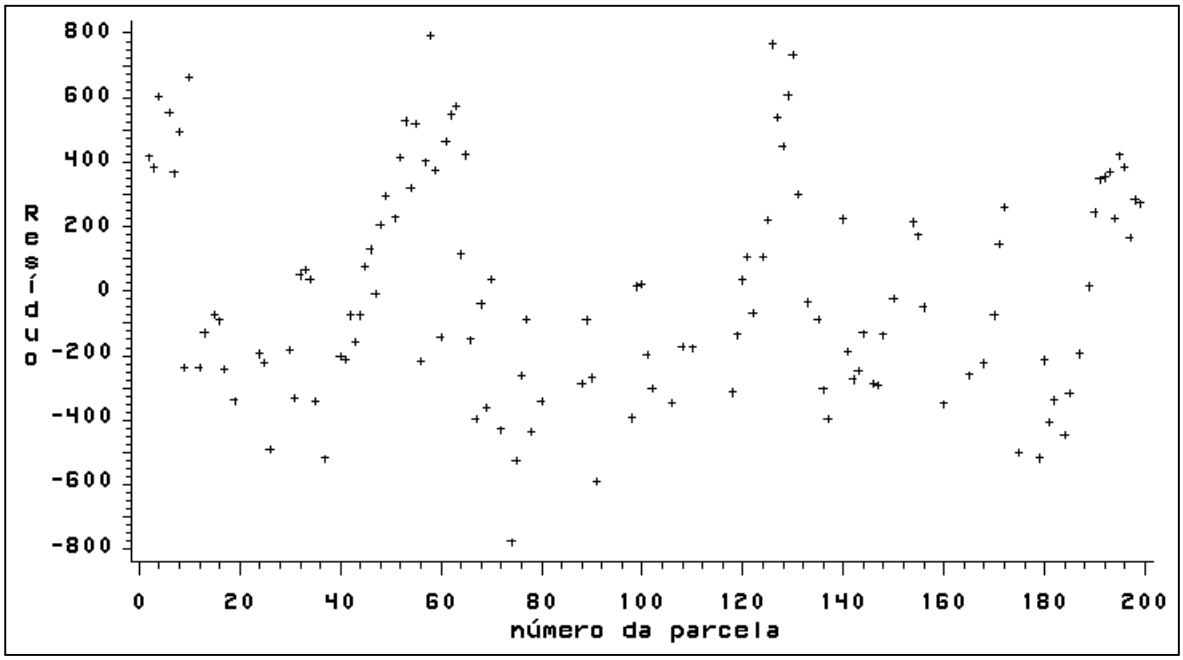

Figura 5.1. Resíduos (kg/ha) do ajuste do modelo de blocos aumentados com recuperação da informação entre novos tratamentos, sob erros independentes, tomados em função do número de identificação das respectivas parcelas no campo experimental.

Na Figura 5.2 confirma-se que os resíduos da análise sob $\mathbf{R}=\mathbf{I}\sigma_e^2$ não se distribuem de forma aleatória no campo experimental. Mas, pelo contrário, há uma tendência nítida de os maiores valores $\hat{e}_{ijk}$ concentrarem-se na parte posterior do mapa de superfície, ou seja, estarem associados a parcelas cujos valores da coordenada X são baixos. Isto determinou, por acréscimo, um gradiente predominante no sentido da largura das parcelas (**COORDX**). Considerando-se que os

Tabela 5.2. Autocorrelações ($\hat{\rho}$) de primeira à décima ordem e estatísticas de Durbin-Watson associadas (d), com as respectivas probabilidades de erro tipo I (H_0: $\rho = 0$), para os resíduos do ajuste de um modelo de blocos aumentados assumindo erros independentes.

Ordem	$\hat{\rho}$	d	Prob<d *	Ordem	$\hat{\rho}$	d	Prob<d
1	0,64125	0,7003	0,0001	6	0,16385	1,5386	0,0143
2	0,50359	0,9599	0,0001	7	0,09528	1,6624	0,0839
3	0,42519	1,0895	0,0001	8	-0,00189	1,8176	0,3440
4	0,38857	1,1313	0,0001	9	-0,06130	1,9241	0,6153
5	0,29623	1,2942	0,0001	10	-0,10327	2,0027	0,7964

* - probabilidade marginal do teste, neste caso, unilateral à esquerda.

blocos foram construídos neste mesmo sentido, constata-se que esta orientação possivelmente não tenha sido a mais adequada, sobretudo à luz da análise de variância usual para os delineamentos em blocos. Dadas as características da superfície de resíduos, a qual fornece uma estimativa do ensaio de uniformidade subjacente ao experimento, é razoável supor que uma blocagem no sentido do comprimento das parcelas teria sido mais efetiva no controle da variação local.

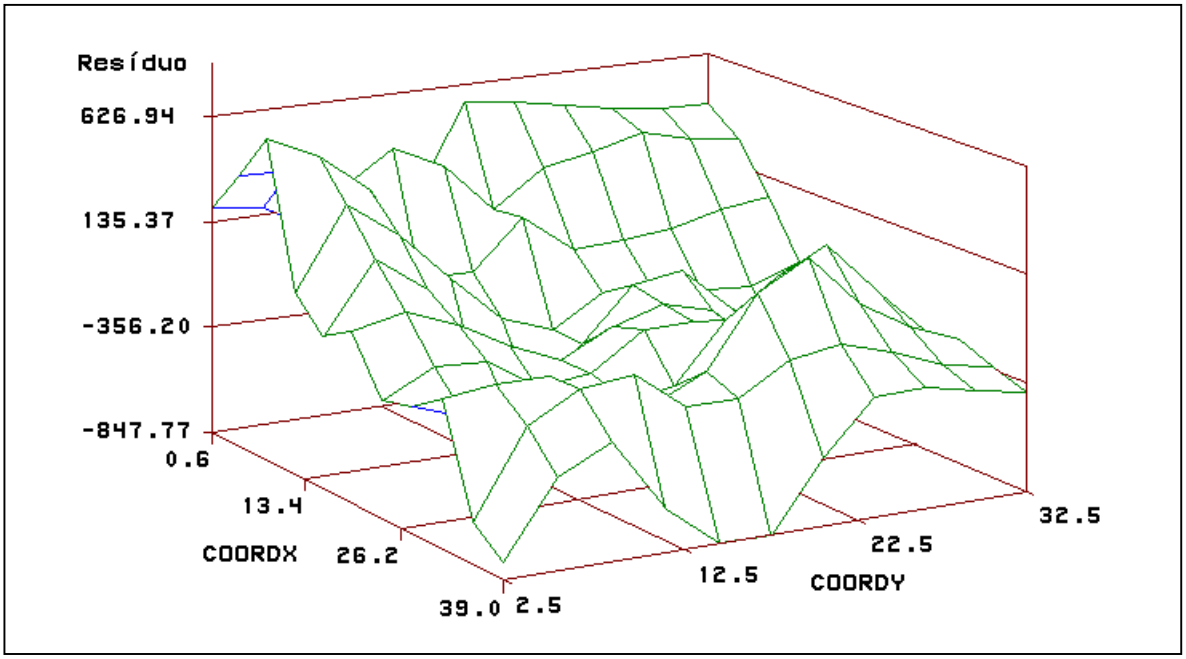

Figura 5.2. Resíduos (kg/ha) do ajuste do modelo de blocos aumentados com recuperação da informação entre novos tratamentos, sob erros independentes, tomados em função das coordenadas, em metros, dos centros das respectivas parcelas (**COORDX** e **COORDY**).

Esse tipo de informação é útil para experimentações futuras na mesma área de terreno e representa, desde já, um mérito da abordagem analítica espacial: o diagnóstico de problemas relacionados ao controle local utilizado. Isto abre a perspectiva da aplicação de um "controle local" *a posteriori*, seja através de uma blocagem alternativa (*post-blocking*), seja pela inclusão de covariáveis no modelo. Este tema, embora de grande interesse prático, não foi objeto deste estudo,

mesmo porque sua aplicação pura e simples pode comprometer a validade dos testes estatísticos (Federer & Schlottfeldt, 1954). Detalhes teóricos e aplicados sobre este assunto podem também ser buscados em Patterson & Hunter (1983), Seraphin (1992) e Federer (1998).

Por último, a Figura 5.3 ilustra o variograma resultante de semivariâncias obtidas para distâncias inferiores a 30m. A configuração dos pontos (variograma amostral) também é típica dos processos estocásticos com dependência espacial, isto é, com variabilidade decrescente à medida que a distância diminui. A representação ainda demostra que a partir de cerca de 20m (alcance prático da correlação espacial), a variabilidade estabiliza-se entre 120000 e 140000 $(kg/ha)^2$. Este patamar representa a variabilidade própria dos resíduos entre parcelas independentes, ou ainda, a covariância residual entre parcelas separadas por uma distância igual ou superior ao alcance. A existência do variograma crescente e com patamar, por sua vez, é um indicativo de que a hipótese de estacionariedade requerida foi satisfeita (Vieira, 2000).

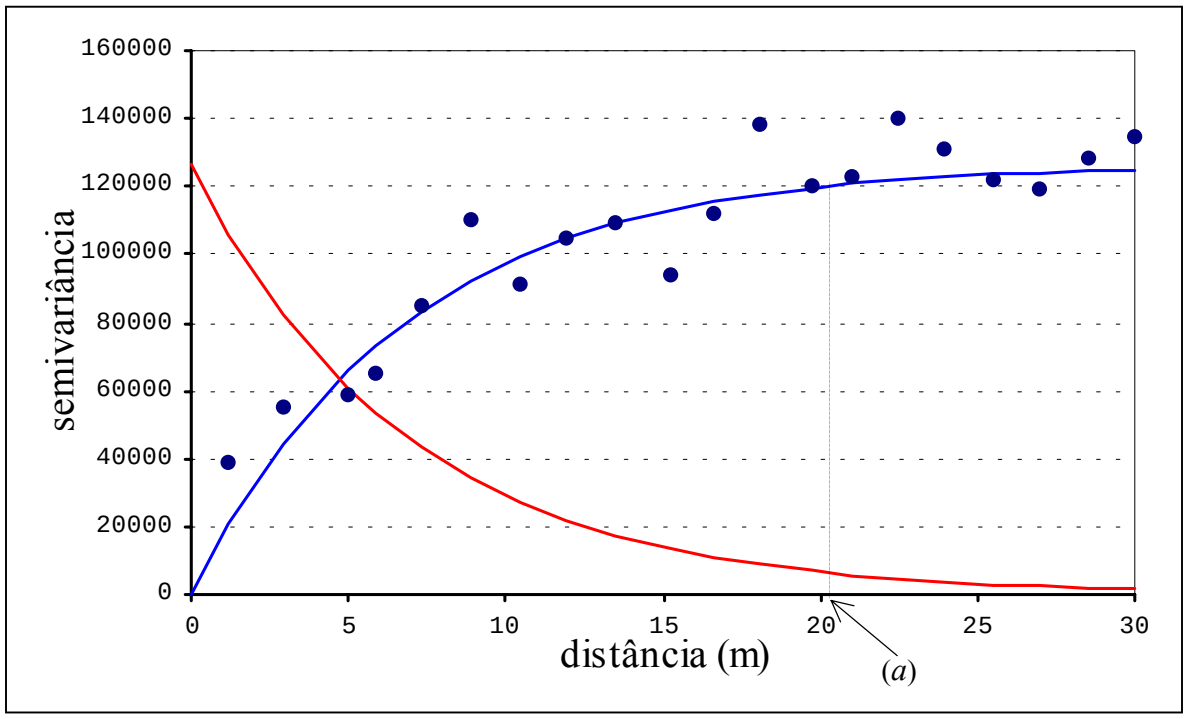

Figura 5.3. Variograma amostral (pontos) dos resíduos da análise de blocos aumentados com recuperação da informação entre tratamentos adicionais, sob erros independentes, e seu respectivo ajuste (linha azul) pelo modelo exponencial de semivariâncias, com parâmetros: a=20,4 m e σ^2=126450 $(kg/ha)^2$ (a linha vermelha ilustra a respectiva função de autocovariância espacial).

Diante disso, sob a suposição de isotropia, a função contínua que melhor se ajustou aos pontos do variograma foi o modelo exponencial de semivariâncias: $S(h) = \sigma^2 [1 - exp(\frac{-3h}{a})]$, com parâmetros σ^2 =126450 $(kg/ha)^2$ e a=20,4 m (Figura 5.3). Por conseguinte, a relação $C(h)=\sigma^2 - S(h)$

determinou a função de autocovariância: $C(h) = 126450\ exp(\frac{-h}{6,8})$. Esta, por sua vez, definiu a matriz de (co)variâncias residuais $\mathbf{R}_{(n)}=\mathbf{\Sigma}$, cujos elementos da diagonal principal foram todos iguais a 126450 e os fora desta diagonal iguais a $126450\ exp(\frac{-h}{6,8})$, sendo *h* a distância que separa cada duas parcelas identificadas por linha e coluna da matriz a que se refere cada elemento. Dessa forma, ficou então caracterizada a covariância espacial inerente ao experimento sob estudo. As implicações do uso ou não desta informação nos resultados da análise estatística são avaliadas na próxima seção.

3.2. Comparação dos modelos de análise espacial e não espacial

A presente discussão basear-se-á nos resultados emitidos pelo *PROC MIXED* do *SAS*, os quais decorreram da execução das rotinas computacionais listadas anteriormente (2.2.2 e 2.2.4). As tabelas aqui apresentadas foram, portanto, organizadas conforme as listagens (saídas) deste procedimento estatístico-computacional.

A princípio, é conveniente avaliar as informações relacionadas à adequação dos modelos de análise (Tabela 5.3). Desde já, faz-se necessário mencionar que, modelos com estrutura de covariância mais complexa (com maior número de parâmetros) sempre tendem a exibir uma melhor qualidade de ajustamento do que os modelos de estrutura mais simples. Se, contudo, esta melhoria não for significativamente importante, demonstra-se que a sobreparametrização do modelo não se justifica. Por isso, devem ser adotados critérios comparativos que, de alguma forma, penalizem os modelos mais parametrizados, a exemplo de *AIC* e *BIC*. No presente caso, observa-se que o modelo espacial apresentou os valores mais elevados para ambos os critérios (*Akaike's Information Criterion* e *Schwarz's Bayesian Criterion*), indicando que a estrutura de covariância espacial ($\mathbf{R}=\mathbf{\Sigma}$) garantiu ao respectivo modelo um ajuste de qualidade superior em relação ao modelo de erros inde-

Tabela 5.3. Informações sobre o ajustamento dos modelos de análise espacial (**RΣ**) e não espacial (**RI**) de um delineamento em blocos aumentados, via *PROC MIXED* do *SAS*, para dados de produtividade de grãos (kg/ha) em soja (ensaio Areão-1994/95, ESALQ-USP).

Descrição	Valor (RI)	Valor (RΣ)
Observations	127,000	127,000
Res Log Likelihood ($l_{REML}(\mathbf{G},\mathbf{R})$)	-881,125	-838,975
Akaike's Information Criterion	-883,125	-841,975
Schwarz's Bayesian Criterion	-885,896	-846,131
-2 Res Log Likelihood ($-2l_{REML}(\mathbf{G},\mathbf{R})$)	1762,251	1677,950
LRT Chi_Sqr (df=1) ($\chi^2_{(1)}$)	-	84,3008 (prob<0,0001)

pendentes ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$). Além disso, esta superioridade foi estatisticamente comprovada pelo teste da razão de verossimilhança (*LRT Chi_Sqr*: $\chi^2_{(1)}$=84,3**), demonstrando, portanto, uma melhor adequação do modelo espacial ao conjunto de dados.

Num segundo momento, é importante avaliar as saídas relacionadas aos componentes de variância (Tabela 5.4). Observa-se que, embora o valor da variância genética (`NEW*GENOT`) tenha sido forçosamente mantido nas duas análises, o erro padrão assintótico listado para esta estimativa reduziu-se consideravelmente em favor da análise espacial. Nota-se também que este valor ($\hat{\sigma}_g^2$=17242,02) não foi estatisticamente significativo na primeira análise (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$), enquanto o foi, em nível de 10% de probabilidade, na segunda análise ($\mathbf{R}=\Sigma$). Isto demonstra que o modelo espacial teve um maior poder para detectar a variabilidade genotípica do que o de observações espacialmente independentes; ou ainda, o modelo espacial mostrou uma maior capacidade de discriminação dos genótipos. A redução no erro assintótico de estimação também ocorreu, favoravelmente à análise espacial, para a variância do erro (σ_e^2 *vs.* σ^2). Por último, vale observar a significância estatística (5% de probabilidade) do parâmetro de covariância *a*, o que confirma a correlação espacial de alcance significativo reportada no item 3.1.

Tabela 5.4. Estimativas dos componentes de (co)variância com os modelos de análise espacial e não espacial de um delineamento em blocos aumentados, respectivos erros padrão assintótico e teste de significância (**Z**) para um valor nulo do parâmetro; dados de produtividade de grãos (kg/ha) em soja (ensaio Areão-1994/95, ESALQ-USP).

Modelo	Parâmetro	Estimativa	Erro padrão	Z	Pr > |Z|
Não Espacial ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$)	NEW*GENOT (σ_g^2)	17242,02	61420,81	0,28	0,7789
	Residual (σ_e^2)	136349,22	57489,82	2,37	0,0177
Espacial ($\mathbf{R}=\Sigma$)	NEW*GENOT (σ_g^2)	17242,02	9822,62	1,76	0,0792
	Variance (σ^2)	126450,00	31712,80	3,99	0,0001
	SP(EXP) (a)	6,80	3,36	2,02	0,0429

Com relação aos testes estatísticos relacionados aos efeitos no modelo, o *PROC MIXED* lista automaticamente apenas os de efeitos fixos (**BLOCO** e **C**), cujos resultados estão na Tabela 5.5. Desdobramentos nestes efeitos e testes de hipóteses associadas aos efeitos aleatórios devem, por isso, ter suas matrizes de contrastes construídas manualmente através do comando '**contrast**' (itens 2.2.2 e 2.2.4). No caso dos efeitos aleatórios, para economizar linhas de programação naquelas rotinas computacionais e tendo em vista apenas o propósito de ilustração, apresentam-se aqui somente os contrastes entre três progênies arbitrariamente escolhidas: USP 93-2007 (**PROG1**),

USP 93-2027 (**PROG3**) e USP 93-2191 (**PROG24**). Os testes correspondentes a esses contrastes são também mostrados na Tabela 5.5.

Tabela 5.5. Testes sobre os efeitos fixos e alguns efeitos aleatórios obtidos dos modelos de análise espacial e não espacial, num delineamento de blocos aumentados (dados de produtividade de grãos, em kg/ha, ensaio de competição de linhagens de soja – Areão: 1994/95, ESALQ-USP).

F. V.	NDF*	Análise não espacial			Análise espacial		
		DDF*	F	Pr > F	DDF*	F	Pr > F
BLOCOS	3	115,0	7,57	0,0001	40,1	1,64	0,1945
C (Populações fixas)	5	12,0	2,71	0,0727	31,9	3,95	0,0067
TESTEMUNHAS	4	11,4	0,93	0,4793	31,5	1,79	0,1547
TESTs vs PROGs	1	15,5	9,89	0,0065	33,3	10,28	0,0030
PROG1 vs PROG3	1	0,20	0,00	0,9859	23,7	0,46	0,5064
PROG1 vs PROG24	1	0,20	0,01	0,9712	25,8	6,00	0,0214
PROG3 vs PROG24	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.

No que se refere aos tratamentos, primeiramente observa-se que a variação em '**C**' (fonte relacionada às seis populações fixas) não foi significativa na primeira análise (a 5% de probabilidade). Mas, atingiu alta significância estatística (prob<0,01) na análise espacial. Quanto aos contrastes obtidos de desdobramentos nesta fonte de variação, nota-se uma elevação nos valores de F, também em favor da análise espacial, tanto na detecção de diferenças entre 'Testemunhas', como na diferenciação entre os grupos 'Testemunhas' e 'Progênies'. No caso das testemunhas isto refletiu numa redução nunca inferior a 12% no erro padrão das médias destes tratamentos. Já na comparação das três progênies entre si (duas a duas), a vantagem da análise espacial foi ainda maior. Enquanto a análise tradicional (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$) não captou diferença alguma entre estes genótipos (prob>0,90), a análise espacial apontou dois dos três contrastes como bastante significativos (prob<0,025). Em síntese, o conjunto desses resultados reflete a maior capacidade de discriminação genotípica do modelo espacial em comparação ao modelo não espacial, e, em última análise, nada menos do que o aproveitamento de um ensaio estatisticamente comprometido à luz da abordagem tradicional.

Este fato pode ainda ser confirmado inspecionando-se os preditores (*EBLUP*'s) dos efeitos genotípicos individuais das progênies (Tabela 5.6, em Anexos). Enquanto na análise tradicional os preditores variaram entre *–98,2* e *100,5*, resultando numa amplitude genotípica de quase *200* kg/ha; na análise espacial esta amplitude foi superior a *500* kg/ha (valores entre *–337* e *200,5*). Isso significa que a diferenciação entre os novos genótipos, em termos destes preditores, aumentou em mais de 150%. Os menores erros padrão associados aos preditores da análise espacial também confirmam a sua melhor discriminação genotípica apontada nos testes de contrastes de médias de

progênies. Enquanto na primeira análise o erro médio dos preditores foi de *124* kg/ha, na análise espacial este erro caiu para *107* kg/ha (redução de 14%). Conclui-se, portanto, que os preditores genotípicos obtidos da abordagem espacial foram consideravelmente de maior precisão.

Finalmente, outro aspecto de grande interesse aplicado refere-se ao ordenamento dos genótipos pelas duas abordagens analíticas (Tabela 5.6). Considerando-se uma seleção hipotética de 25% das linhagens mais produtivas (28 em 110 genótipos), observa-se uma coincidência de apenas 46% entre as duas seleções. Além disso, entre os genótipos hipoteticamente selecionados pela análise tradicional, pelo menos 30% deles ocuparam más posições de classificação (acima da 50ª) na análise espacial. Alguns exemplos são as linhagens: USP 93-2048, USP 93-2393, USP 93-2153 e USP 93-2198. Por outro lado, as linhagens USP 93-2027, USP 93-2693, USP 93-2785 e USP 93-2790, classificadas entre as dez mais produtivas na análise espacial, seriam todas descartadas na análise alternativa (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$).

Diante das discordâncias nestas seleções, outra curiosidade natural é verificar a localização das parcelas em que foram testados os melhores genótipos, segundo as diferentes abordagens. Neste sentido, a Figura 5.4 ilustra os efeitos do ajustamento espacial sobre a seleção. Quando o modelo não espacial foi utilizado, os genótipos selecionados vieram exclusivamente da faixa lateral esquerda do campo experimental (parcelas com baixos valores de **COORDX**), provavelmente a sua área mais fértil. Porém, quando o ajuste espacial foi levado em conta, os genótipos selecionados vieram de parcelas espalhadas por todo o experimento. A maior predominância de genótipos ainda vindos daquela faixa pode ser explicada por efeitos de fertilidade remanescentes e/ou pela prefe-

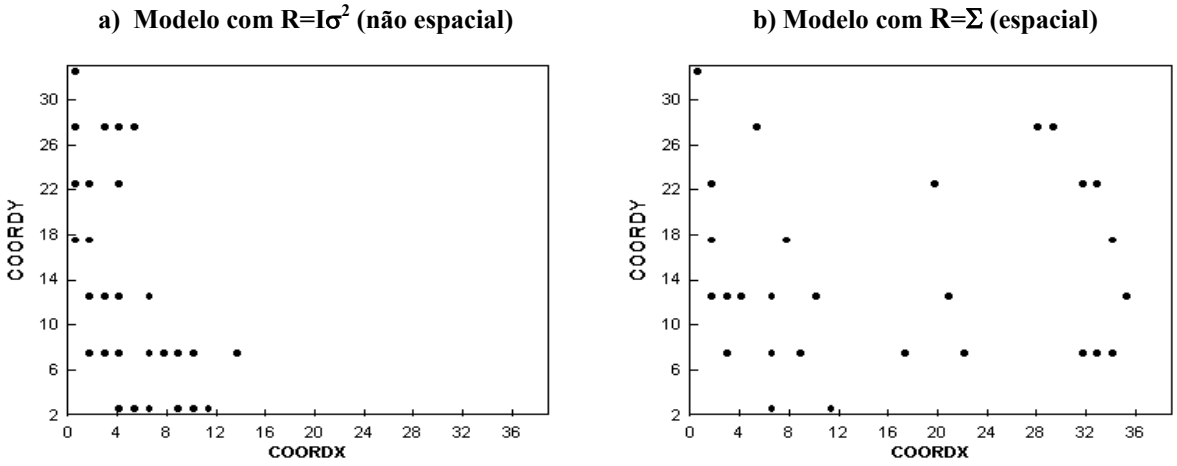

Figura 5.4. Localização no campo experimental das parcelas em que foram testadas as linhagens mais produtivas (25%), entre 110 novos genótipos (não replicados), de acordo com dois modelos de análise estatística (**a** e **b**) (ensaio de competição de linhagens de soja delineado em blocos aumentados: Areão-1994/95, ESALQ-USP).

rência em alocar genótipos de um mesmo parental lado a lado. De qualquer forma, o que se espera para a realidade experimental é algo próximo do que se visualiza na parte (b) da Figura 5.4 e não o que se vê na parte (a). Resultados semelhantes foram também obtidos por Besag & Kempton (1986) e Cullis *et al*. (1989), os quais são objetos de ilustração em Kempton & Gleeson (1997).

Enfim, considerando-se que a causa da divergência nas duas seleções foi o ajustamento genotípico para efeitos de posição (ambientais), conclui-se que, em situações similares, o uso da análise espacial pode garantir um maior sucesso ao programa de melhoramento. Isto não significa, portanto, uma recomendação incondicional desse tipo de análise. Os resultados são, antes de tudo, um alerta para os usuários de blocos aumentados sobre as implicações de uma possível correlação espacial entre parcelas. Aqui, isso está dirigido particularmente aos melhoristas de plantas, que se vêem, muitas vezes, obrigados a adotar parcelas de pequeno tamanho e/ou testemunhas ou grupos de genótipos alocados sistematicamente. A recomendação básica que se faz, nestes casos, é uma atenção redobrada à suposição clássica de independência; e, nas situações em que esta for violada, recomenda-se, então, o uso de métodos analíticos menos restritivos tal como a abordagem espacial de campo aleatório. Só assim a análise estatística se constituirá numa ferramenta eficaz do melhorista para a seleção de genótipos realmente superiores; haja vista a busca de testes de hipóteses, estimativas e predições cada vez mais coerentes com a realidade.

4. CONCLUSÕES

Os resultados da presente investigação permitiram concluir, inicialmente, que o enfoque estatístico espacial fornece um diagnóstico de problemas relacionados ao controle local utilizado nos experimentos. Assim, possibilita orientar o pesquisador para a tomada de decisões imediatas ou futuras no sentido de um controle mais efetivo da variabilidade local. Pôde-se concluir também que, nos ensaios genéticos com parcelas de pequeno tamanho e normalmente sem bordadura, a correlação espacial entre observações pode atingir magnitude e alcance altamente significativos. Isso compromete a validade e a eficiência dos métodos estatísticos que assumem erros independentes ($\mathbf{R}=\mathbf{I}\,\sigma_e^2$), exigindo, portanto, uma atenção especial à validade das suposições da análise de variância clássica.

Na presença desse tipo de autocorrelação, a qualidade de ajustamento de um modelo de análise espacial pode ser bastante superior à dos modelos tradicionais (sob $\mathbf{R}=\mathbf{I}\,\sigma_e^2$). Particularmente no caso dos ensaios de melhoramento de plantas, isto reflete no aumento do poder dos testes estatísticos, no alargamento da amplitude das predições genotípicas, bem como na redução dos

erros padrão de médias e de preditores genotípicos. Sendo assim, a decisão entre adotar ou não uma análise estatística espacial pode simplesmente determinar o aproveitamento do ensaio, enquanto critério para a discriminação entre os tratamentos genéticos. Ademais, sob autocorrelação, os genótipos selecionados pelas duas abordagens (espacial e não espacial) podem diferir substancialmente entre si em decorrência do ajuste para os efeitos de posição. Considerando-se que tais efeitos são de natureza puramente ambiental, pode-se concluir que, nestas condições, a seleção de genótipos baseada na análise espacial é seguramente superior.