ISSN 1517-536X
Novembro, 2001

Empresa Brasileira de Pesquisa Agropecuária
Centro Nacional de Pesquisa de Florestas
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 65

Análise **G**enética de **D**ados com **D**ependência **E**spacial e **T**emporal no **M**elhoramento de **P**lantas **P**erenes via **M**odelos **G**eoestatísticos e de **S**éries **T**emporais **E**mpregando REML/BLUP ao **N**ível **I**ndividual

Marcos Deon Vilela de Resende
José Alfredo Sturion

Colombo, PR
2001

Exemplares desta publicação podem ser adquiridos na:

Embrapa Florestas
Estrada da Ribeira km 111 - CP 319
83411-000 - Colombo, PR - Brasil
Fone: (41) 666-1313
Fax: (41) 666-1276
Home page: www.cnpf.embrapa.br
E-mail (sac): sac@cnpf.embrapa.br

Supervisor editorial: Moacir José Sales Medrado
Revisor de texto: Elly Claire Jansson Lopes
Normalização bibliográfica: Lidia Woronkoff
Tratamento de ilustrações: Cleide Fernandes de Oliveira
Foto(s) da capa: arquivos da Embrapa
Editoração eletrônica: Marta de Fátima Vencato

1ª edição
1ª impressão (2001): 500 exemplares

CIP-Brasil. Catalogação na publicação.
Embrapa Florestas

Marcos Deon Vilela de Resende

Análise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de séries temporais empregando REML/BLUP ao nível individual / Marcos Deon Vilela de Resende, José Alfredo Sturion. Colombo: Embrapa Florestas, 2001. 79 p. (Embrapa Florestas. Documentos, 65).

Inclui bibliografia
ISSN 1517-536X

1. Planta perene - Melhoramento genético. 2. Espécie florestal - Melhoramento genético. 3. Experimentação - Modelo geoestatístico. 4. Experimentação - Série Temporal. 5. REML/BLUP. I. Título. II. Série.

CDD 582.16

Sumário

Análise Genética de Dados com Dependência Espacial e Temporal no Melhoramento de Plantas Perenes Via Modelos Geoestatísticos e de Séries Temporais Empregando REML/BLUP ao Nível Individual

Marcos Deon Vilela de Resende
José Alfredo Sturion

1. Introdução

A análise tradicional de experimentos de campo parte do princípio de que todas as observações tomadas em posições adjacentes (em plantas ou parcelas vizinhas) são não correlacionadas. Desta forma, a matriz de covariância residual ou dos erros é modelada como $R = I\sigma_e^2$, ou seja, os erros são assumidos como independentes. Também, a posição dos tratamentos no campo, ou seja, a distribuição espacial dos mesmos é ignorada. Segundo Steel & Torrie (1980), a casualização concorre para a neutralização dos efeitos da correlação espacial e, portanto, para a geração de uma análise de variância fidedigna.

Entretanto, conforme Grondona et al. (1996), embora a teoria da casualização enfatize a neutralização da correlação espacial, tal neutralização é mais eficiente quando se usam modelos espaciais. Também as formas de controle local baseadas em blocagem podem ser ineficientes para tratar de problemas de gradientes ambientais e mesmo os blocos incompletos não permitem uma avaliação completa dos efeitos espaciais. Além disso, a blocagem é realizada antes da implantação dos experimentos, de forma que percebe-se muitas vezes (principalmente em espécies perenes), por ocasião da coleta dos dados experimentais, a presença de manchas ou gradientes ambientais dentro dos experimentos, os quais não foram considerados adequadamente pelos blocos delineados "a priori". Nesta situação, somente as técnicas de análise espacial, permitem contornar a questão e propiciar uma seleção acurada, através de blocagem "a posteriori" ou através da flexibilização da matriz R baseados nos próprios dados experimentais.

A autocorrelação espacial advinda de observações com dependência espacial compromete as suposições de homogeneidade dentro de blocos e de ausência de interação genótipos x blocos, suposições estas essenciais para se obter eficiência no delineamento experimental e boa capacidade de teste. Assim, a constatação de ocorrência de autocorrelação espacial e a consideração de uma estrutura de erros correlacionados com $R = \Sigma$, em que Σ refere-se a uma matriz não diagonal, pode permitir a obtenção de estimativas e predições mais precisas, aumentando assim, a eficiência da análise estatística.

A consideração deste tipo de estrutura ($R = \Sigma$) é também relevante na análise de medidas repetidas tomadas sobre um mesmo indivíduo em plantas perenes, sendo uma opção aos modelos de repetibilidade, multivariado e de regressão aleatória. A autocorrelação espacial ou serial é positiva e aumenta com a diminuição da distância entre as observações. Dessa forma, modelos que consideram a correlação entre medidas repetidas como sendo a mesma para quaisquer pares ou combinações de idades ou safras podem não ser os mais adequados. Assim sendo, os modelos de análise espacial podem ser vantajosos também para a análise de medidas repetidas, ou seja, para análise de dados com dependência temporal.

O presente artigo tem como objetivos discorrer sobre a aplicação de métodos geoestatísticos e de análise de séries temporais na análise de experimentos de campo delineados no contexto do melhoramento de plantas perenes. Ênfase é dada à análise de dados com dependência espacial, mas extensões à análise de dados com dependência temporal são triviais.

O objetivo final do melhoramento em benefício da produção vegetal refere-se à maximização da expressão fenotípica do caráter de interesse no campo do produtor rural. Uma vez que a expressão fenotípica é função do genótipo, ambiente e interação genótipo x ambiente, torna-se relevante o estudo destes três fatores. Neste contexto, a predição dos valores genéticos é de fundamental importância, mas o estudo dos fatores ambientais é de igual relevância.

Visando a separação dos efeitos genéticos e ambientais são utilizadas técnicas de genética quantitativa aplicadas sobre dados obtidos de experimentação a campo. Uma tentativa de refinamento no estudo dos genótipos refere-se à utilização de técnicas de biologia molecular notadamente a análise QTL's (locos

controladores de características quantitativas) marcados. Um refinamento no estudo dos efeitos ambientais refere-se ao emprego de técnicas de análise espacial (via modelos geoestatísticos e de séries temporais), as quais permitem uma análise mais detalhada da variabilidade espacial dos solos, permitindo uma melhor estratificação ambiental (identificação de estratos mais homogêneos e de gradientes ambientais) e uma comparação mais efetiva dos genótipos através do uso de uma estrutura de erros espacialmente dependentes e correlacionados.

2. Estratégias de Controle Local e de Consideração da Correlação Espacial para aumento da Precisão Experimental

A variabilidade ou heterogeneidade espacial associada à fertilidade e estrutura do solo, umidade, interceptação de luz e outros fatores ambientais contribuem para o aumento da variação residual. Assim, é importante controlar, por delineamento ou por análise, a variação residual (Grondona et al., 1996).

Um delineamento experimental adequado deve obedecer aos princípios fundamentais da experimentação: repetição, casualização e controle local. A importância do número de repetições é capital, significando que, com baixo número de repetições, até a casualização é prejudicada ou comprometida. Como controle local deve ser enfatizada a homogeneidade dentro de estratos ou blocos, sendo, em princípio, recomendados os delineamentos em blocos casualizados e látice. A casualização e a repetição é que propiciam uma comparação não viciada dos tratamentos, ao passo que o controle local e a repetição permitem reduzir o erro experimental médio. Um erro experimental menor permite inferir como significativa uma diferença real pequena entre médias de tratamentos (Steel & Torrie, 1980).

O delineamento em quadrado latino, provavelmente, é o que propicia melhor controle local, visto que permite controlar a heterogeneidade ambiental em duas direções, no sentido das linhas e das colunas. Entretanto, tal delineamento não tem sido recomendado para os trabalhos de melhoramento (Ramalho et al., 2000) ou na experimentação em geral (Pimentel Gomes, 1987), devido à restrição do número de repetições ter que ser igual ao número de tratamentos ou

progênies. Dessa forma, não se tem relatos de sua utilização no melhoramento. No entanto, com o advento da utilização de parcelas de uma planta no melhoramento de plantas perenes, tal delineamento passa a ter grande potencial de utilização. Como se utilizam em torno de 60 plantas por progênie (60 repetições de uma planta), quadrados latinos de 60 x 60 com 60 progênies poderiam ser perfeitamente utilizados, em associação com o procedimento BLUP. No caso, os dados seriam corrigidos para dois gradientes ambientais (linhas e colunas), pelo método BLUP ou do índice multi-efeitos. Segundo Panse e Sukhatme (1963), quando existem tendências simultâneas de variações em fertilidade em duas direções em ângulos retos (que equivale a uma tendência diagonal em fertilidade), é provável que o quadrado latino seja mais eficiente que o delineamento em blocos. O delineamento em quadrado latino é também recomendado quando não se conhece a priori os gradientes de fertilidade.

Os delineamentos de blocos incompletos (látice, por exemplo) são especialmente indicados na situação de grande número de tratamentos e alta variabilidade ambiental (quantificada pelo b de Smith, por exemplo) na área experimental. No melhoramento de espécies florestais, o delineamento de blocos casualizados tem sido o mais utilizado na América do Norte (Fu et al., 1998), os blocos incompletos tem sido os mais utilizados na Austrália, África do Sul e Ásia (Williams & Matheson, 1994) e no Brasil ambos os tipos de delineamentos tem sido utilizados.

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

Outra classe de delineamentos que tem sido utilizada de maneira crescente nos últimos anos são os blocos aumentados de Federer (Federer, 1958, 1998; Wolfinger et al., 1997), os quais, por construção são desbalanceados. Os delineamentos de blocos aumentados, blocos incompletos balanceados e parcialmente balanceados não são ortogonais. Neste caso, o uso destes delineamentos para seleção, conduz, via análise intrablocos, a médias de tratamentos viciadas, mesmo quando a sobrevivência for 100% (Resende e

Fernandes, 2000). Não significa isto, que estes delineamentos não devem ser usados mas, que sejam usados em associação com o procedimento BLUP, o qual ajusta as médias para as estimativas BLUE dos efeitos fixos ou efeitos ambientais identificáveis.

Entretanto, os delineamentos em blocos baseiam-se na premissa de conhecimento a priori da heterogeneidade da área experimental de forma que seja possível alocar todas as parcelas (tratamentos) em blocos homogêneos (Lotode, 1971). Caso esta heterogeneidade não seja conhecida a priori, a delimitação dos blocos tornar-se-á arbitrária, fato que poderá implicar forte heterogeneidade dentro de blocos. Com base no exposto, uma alternativa é a alocação aleatória de parcelas de uma planta no campo experimental e posterior controle da heterogeneidade ambiental, empregando-se métodos tais como o da análise de covariância associando uma covariável independente à variável estudada (Método de Papadakis, Papadakis (1984)), ou das variáveis regionalizadas ou espaciais (Métodos Geoestatísticos, conforme Lecoustre & Reffye (1986)).

O ajustamento a posteriori para os gradientes ambientais em testes de progênies apresenta um potencial para um aumento significativo da eficiência na estimação de parâmetros genéticos e seleção. Neste contexto, o uso de parcelas de uma planta completamente aleatorizadas (delineamento inteiramente ao acaso), tem sido novamente comum (Lotode & Lachenaud, 1988). Entretanto, Gilmour (2000) adverte para o fato de que a blocagem a posteriori não deve ser baseada apenas na significância estatística de contrastes arbitrários. O experimentador necessita identificar as causas físicas e ambientais que levaram à determinado tipo de blocagem. É importante relatar que o próprio uso do delineamento em quadrado latino pode reduzir a necessidade do uso de técnicas (como a análise espacial) de ajustamento a posteriori.

Outras formas de controle local e aumento da precisão experimental referem-se à análise de covariância ou ajuste de covariável e aos métodos de análise de vizinhança. Os métodos de análise de vizinhança fundamentam-se no princípio de vizinhança e baseiam-se no ajuste de uma covariável associada a observações de parcelas vizinhas, sendo que os dois principais métodos são: (i) médias móveis (MA), baseado na média das parcelas vizinhas; (ii) Papadakis, baseado na média dos erros das parcelas vizinhas. Segundo Pearce (1998), ao se usar o método de Papadakis não se faz necessário usar a blocagem simultaneamente.

Outros procedimentos destinados ao controle da correlação positiva entre observações obtidas de plantas vizinhas são: (i) método das variáveis regionalizadas ou procedimentos geoestatísticos (Cressie, 1993; Martinez, 1994); (ii) métodos de análise de séries temporais, via modelos AR, MA e ARIMA em duas dimensões (Martin, 1990; Cullis & Gleeson, 1991; Gilmour et al., 1997; Gilmour et al., 1998; Cullis et al. 1998); (iii) modelo linear de campo aleatório (Zimmerman & Harville, 1991); (iv) ajuste de superfícies de resposta polinomiais (Federer, 1998). Estes procedimentos são considerados em maior detalhe em tópicos específicos neste artigo.

3. Análise Ambiental de Experimentos e Avaliação da Necessidade da Abordagem Espacial

Na análise de experimentos em genética e melhoramento de plantas conduzidos em um único ambiente, tradicionalmente tem-se enfatizado mais a análise genética do que a ambiental. Embora a seleção deva basear-se em um ordenamento dos valores genéticos dos indivíduos candidatos à seleção, a utilização prática e comprovação do valor real dos materiais genéticos melhorados baseia-se em seus valores fenotípicos, os quais são influenciados pelo ambiente. Isto justifica uma análise mais detalhada dos efeitos ambientais em um experimento.

Tal análise ambiental deve enfatizar pelo menos três fatores: (i) a eficiência do delineamento em termos do controle local; (ii) a variabilidade espacial dentro dos estratos ambientais homogêneos (blocos); (iii) a interação genótipo x ambiente dentro de um mesmo sítio ou fazenda. O fator (i) pode ser estudado com base na significância do teste F de Snedecor associado a fonte de variação blocos na análise de variância e também com base no coeficiente de correlação intraclasse entre parcelas dentro dos blocos (ρ_b). A variabilidade espacial dentro dos blocos pode ser estudada através do coeficiente de correlação intraclasse entre indivíduos de uma mesma parcela, devido ao ambiente comum da parcela (c^2) o qual pode, alternativamente, ser denominado coeficiente de determinação dos efeitos ambientais entre parcelas. Por sua vez, o fator (iii) pode ser investigado com base na correlação genética intraclasse (ρ_g), dos materiais genéticos ao longo das repetições, ou seja, de uma repetição para outra.

A eficiência do delineamento experimental em blocos (DBC), avaliada a partir de uma análise ao nível de médias de parcela é apresentada a seguir, considerando os efeitos de blocos como aleatórios. Considere os seguintes resultados associados à avaliação de c = 62 clones de caju em b = 3 blocos, para o caráter diâmetro da copa.

F.V.	G.L.	Q.M.	Q.M.	E (QM)	F
Bloco	2	Q_1	0,41845	$\sigma^2 + c\sigma_b^2 = \sigma^{2*}[1+(c-1)\rho_b]$	4,14*
Clones	61	Q_2	0,17912	-	-
Resíduo	122	Q_3	0,10106	$\sigma^2 = \sigma^{2*}(1-\rho_b)$	

* significativo ao nível de 5% de probabilidade de erro tipo I.

Tem-se as seguintes estimativas:

$\hat{\sigma}^2 = Q_3 = 0{,}10106$: estimativa da variância residual.

$\hat{\sigma}_b^2 = (Q_1 - Q_3)/c = 0{,}005119$: estimativa da variância entre blocos.

: $\hat{\rho}_b = \hat{b}^2 = \dfrac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}^2} = 0{,}04821$: estimativa da correlação intraclasse entre parcelas dentro dos blocos.

Sendo $\hat{\sigma}^{2*} = \hat{\sigma}_b^2 + \hat{\sigma}^2$, a esperança da variância residual para um delineamento inteiramente ao acaso (DIC) e $1/\hat{\sigma}^{2*}$ *e* $1/\hat{\sigma}^2$ as **quantidades de informação** associadas ao DIC e ao DBC, respectivamente, tem-se que a eficiência do DBC em relação ao DIC é dada pela razão entre as respectivas quantidades de informação, ou seja, por

$$E_{13} = (1/\hat{\sigma}^2)/(1/\sigma^{2*}) = \frac{\hat{\sigma}^{2*}}{\hat{\sigma}^2} = \frac{\hat{\sigma}^2 + \hat{\sigma}_b^2}{\hat{\sigma}^2} = \frac{1}{1-\hat{\rho}_b} = 1{,}0507 .$$

Verifica-se, pela expressão de E_{13}, que a eficiência de um DBC em relação a um DIC é tanto maior quanto maior for o coeficiente de correlação intraclasse ρ_b. No presente exemplo, tal eficiência foi baixa, revelando a existência de heterogeneidade entre blocos na área experimental (a estatística F para blocos mostrou-se significativa), mas de baixa magnitude.

A significância dos efeitos de blocos deve ser analisada em conjunto com o parâmetro c^2. Assim, tem-se quatro situações: (a) F para blocos significativo e c^2 alto; (b) F para blocos significativo e c^2 baixo; (c) F para blocos não significativo e c^2 baixo; (d) F para blocos não significativo e c^2 alto. Um c^2 alto significa alta variabilidade entre parcelas no bloco e um c^2 baixo significa baixa variação entre parcelas no bloco. Considerando o conceito de **capacidade de teste** como a capacidade do experimento propiciar aos materiais genéticos experimentar diferentes condições ambientais, pode-se fazer as inferências apresentadas a seguir.

Na situação (a), pode-se dizer que o delineamento não foi totalmente eficiente mas que a capacidade de teste foi adequada. Isto porque, embora os blocos tenham apresentado diferenças significativas entre eles, uma grande heterogeneidade ambiental dentro dos blocos permaneceu. Em (b) pode-se inferir que o delineamento foi eficiente e a capacidade de teste adequada. Em (c), a inferência é que existe uma grande homogeneidade ambiental na área experimental e, neste caso, qualquer delineamento é eficiente mas não existe uma capacidade de teste adequada, podendo-se incorrer no risco de se selecionar materiais genéticos com pequena **plasticidade fenotípica**. A situação (d), por sua vez, denota que o delineamento não foi eficiente e a capacidade de teste inadequada. Neste último caso, o melhorista deve procurar métodos mais sofisticados de análise, tais como uma análise espacial visando a realização de uma blocagem a posteriori. As causas dos resultados associados à situação (d) podem ser atribuídos a: (i) blocos muito grandes foram alocados, de forma que a variação dentro de blocos tendeu a ser próxima da magnitude da variação entre blocos (a correlação intraclasse entre parcelas dentro de bloco, ρ_b, foi muito baixa); (ii) o gradiente ambiental (de fertilidade, por exemplo) ocorre em vários sentidos. As quatro situações mencionadas encontram-se sintetizadas na Tabela 1.

Tabela 1. Inferências práticas sobre eficiência de delineamento e capacidade de teste, em função das estatísticas F de Snedecor para blocos e coeficiente de determinação dos efeitos de parcela (c^2).

Situação	F	c^2	Eficiência de delineamento	Capacidade de teste
(a)	Significativo	Alto	Não	Sim
(b)	Significativo	Baixo	Sim	Sim
(c)	Não significativo	Baixo	Sim	Não
(d)	Não significativo	Alto	Não	Não

Em resumo, este tipo de análise deve ser realizada e tem-se as seguintes implicações práticas de acordo com a situação a que os resultados remetem:

- situação (a): o melhorista deve utilizar métodos mais sofisticados de análise;
- situação (b): é a situação ideal ao melhorista;
- situação (c): o melhorista deve prever uma maior perda de ganho genético realizado devido à interação genótipo x ambiente;
- situação (d): o melhorista deve lançar mão de métodos mais sofisticados de análise, mas não necessariamente conseguirá uma capacidade de teste adequada.

Nas situações (a) e (d), o melhorista deveria ter usado outro delineamento, como o látice ou o quadrado latino. É importante mencionar que os métodos de blocagem a posteriori tenderão a propiciar maiores eficiências de delineamento e capacidades de teste quando os experimentos forem implantados no delineamento inteiramente casualizado e com uma planta por parcela.

A variabilidade espacial dentro de blocos e a interação genótipo x ambiente dentro de um mesmo sítio podem ser investigados tomando-se por base uma análise de um DBC ao nível de plantas individuais. Considerando o mesmo experimento com clones de caju, com n = 4 plantas por parcela, tem-se os seguintes resultados da análise de variância:

F.V.	G.L.	Q.M.	Q.M.	E (QM)	F
Bloco (B)	6	Q_1	1,6738	-	-
Clones(C)	61	Q_2	0,71648	$\sigma_\delta^2 + n\sigma_e^2 + nb\sigma_g^2$	F=1,7724
Resíduo(BxC)	122	Q_3	0,40424	$\sigma_\delta^2 + n\sigma_e^2$	F* =1,1392*
Dentro de parcela	558	Q_4	0,35484	σ_δ^2	

* não significativo.

As estimativas dos parâmetros de interesse são:

$\hat{\sigma}_\delta^2 = Q_4 = 0{,}35484$: estimativa da variância dentro de parcelas.

$\hat{\sigma}_e^2 = (Q_3 - Q_4)/n = 0{,}01235$: estimativa da variância entre parcelas.

$\hat{\sigma}_g^2 = (Q_2 - Q_3)/nb = 0{,}02602$: estimativa da variância genotípica entre clones.

$\hat{c}^2 = \dfrac{\hat{\sigma}_e^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_e^2 + \hat{\sigma}_g^2} = 0{,}0314$: estimativa do coeficiente de determinação dos efeitos de parcela.

$\hat{h}_g^2 = \dfrac{\hat{\sigma}_g^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_e^2 + \hat{\sigma}_g^2} = 0{,}3932$: estimativa da herdabilidade individual no sentido amplo.

$\hat{\rho}_g = \dfrac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \dfrac{1}{1 + \dfrac{(F^*-1)}{F-1}\dfrac{b}{F^*}} = 0{,}6781$: estimativa da correlação genética intraclasse dos materiais genéticos através das repetições.

Interpretando-se os valores de F para blocos e $\hat{c}^2$, pode-se enquadrar o presente experimento na situação (b), que é a ideal ao melhorista e, portanto, pode-se inferir que a experimentação foi adequada.

Os valores de $\hat{c}^2$ observados em bons experimentos em plantas perenes situam-se em torno de 0,10 (quando a herdabilidade estimada é da ordem de 0,30), ou

seja, 10% da variação fenotípica total dentro do bloco. Assim, para um nível de herdabilidade individual ao redor de 0,30, $\hat{c}^2 \leq 0{,}10$ podem ser classificados como baixos e $\hat{c}^2 > 0{,}10$ podem ser classificados como altos, permitindo assim, alguma inferência sobre a variabilidade espacial dentro dos blocos.

O parâmetro ρ_g é útil na inferência sobre a interação materiais genéticos x blocos, revelando que tanto menor é a interação quanto maior for ρ_g. Por extensão, tal parâmetro permite inferir também sobre a interação genótipo x ambiente dentro do próprio sítio ou fazenda de plantio. Imaginando-se que toda a área de plantio em uma fazenda comporta um grande número de blocos diferentes ($b \to \infty$), o interesse do melhorista é indagar sobre a capacidade da média de um genótipo sobre os b blocos do experimento, correlacionar-se com a média dos mesmos genótipos sobre os $b \to \infty$ blocos ou seja com o valor genotípico real do indivíduo.

Esta correlação é dada por $r_{\hat{g}g\infty} = [r_{ggd}]^{1/2} = \left[\frac{b\rho_g}{1+(b-1)\ \rho_g}\right]^{1/2}$, em que r_{ggd} foi definida no tópico 19.10 como a correlação genética dos materiais genéticos dentro de um sítio ou local e eqüivale a $\hat{r}_{ggd} = \frac{b\hat{\rho}_g}{1+(b-1)\ \hat{\rho}_g} = 0{,}8634$, no presente caso. Neste exemplo, a interação genótipo x blocos não foi significativa e a correlação ρ_g apresentou magnitude moderada, fato que deverá concorrer para uma pequena redução no ganho genético realizado, devido à ocorrência de interação genótipo x ambiente.

Em realidade r_{ggd} é um coeficiente de determinação da média dos materiais genéticos através dos blocos e pode ser usado no cômputo do ganho genético com a seleção em um local, através da expressão:

$$G_s = d_s\ r_{ggd}\ h^2_{mc}$$

$$= K\ r_{ggd}\ h^2_{mc}\ \sigma_{Fm}$$

A acurácia seletiva, no caso, é dada por $r_{g\hat{g}} = \left[r_{ggd}\ h^2_{mc}\right]^{1/2}$.

No presente exemplo, selecionando-se 10% (k = 1,755) dos clones tem-se:

$$\hat{\sigma}^2_{Fm} = \frac{Q_2}{nb} = \frac{0{,}71648}{12} = 0{,}0597$$

$$h^2_{mc} = 1-(1/F) = 1-(1/1{,}7724) = 0{,}4358$$

$\hat{G}_{snc} = 1{,}755 \ . \ 0{,}4358 \ . \ (0{,}0597)^{1/2} = 0{,}1869$: estimativa do ganho genético não corrigido para a perda devida à interação genótipo x ambiente;

$\hat{G}_{sc} = 1{,}755 \ . \ 0{,}8634 . \ 0{,}4358 \ . \ (0{,}0597)^{1/2} = 0{,}1614$: estimativa do ganho genético corrigido para a perda devida à interação genótipo x ambiente;

$\hat{r}_{\hat{g}g} = [0{,}8634 \ . \ 0{,}4358]^{1/2} = 0{,}6134$: estimativa da acurácia seletiva.

Assim, o ganho corrigido equivale a 86,34% do ganho não corrigido e reflete a perda devida à interação genótipo x ambiente dentro de locais.

Simmonds (1989) relata, para seringueira, que a produção dos clones em plantios comerciais é sempre inferior a produção dos mesmos nos ensaios. Este fato pode ser atribuído à perdas devidas à interação genótipo x ambiente. As metodologias aqui relatadas podem contribuir para uma melhor inferência sobre a produtividade comercial esperada dos clones. As estimativas de componentes de variância necessárias na aplicação destes métodos não necessitam ser estimadas via análise de variância, podendo ser estimadas também via metodologia de modelos mistos. Os procedimentos apresentados neste tópico são igualmente aplicáveis aos testes de progênies e não apenas aos testes clonais.

Em resumo, previamente à realização de análises espaciais, deve-se procurar verificar a existência ou não de uma estrutura de dependência espacial nos experimentos, ou seja, deve-se diagnosticar a presença de variabilidade espacial. E isto pode ser verificado de maneira simples com base nos parâmetros c^2 e ρ_g. Valores altos de c^2 indicam alta variabilidade entre parcelas dentro de blocos e alta correlação ambiental entre observações dentro de parcela. Por sua vez, valores baixos de ρ_g indicam que os tratamentos são ordenados

diferentemente de um bloco para outro, fato que pode ser devido a um gradiente de fertilidade, segundo Eisenberg et al. (1996).

Outros testes estatísticos simples destinados à avaliação da autocorrelação ou correlação serial de resíduos em análise de séries temporais ou de regressão, como o teste de Durbin-Watson, podem também ser aplicados na avaliação da dependência espacial. O teste de Durbin-Watson é dado por (Morettin & Toloi, 1987):

$dw = \frac{\sum_{1=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{1=1}^{n} (\hat{e}_i^2)}$, em que 1 = 1, 2,....,n refere-se à ordem da posição da

observação vinculada ao erro $\hat{e}_i$. No caso, $\hat{e}_i$ e $\hat{e}_{i-1}$ são erros de observações em que as posições são adjacentes, ou seja, de observações com vizinhança de primeira ordem.

A estatística dw apresenta relação direta com a autocorrelação de primeira ordem (ρ), dada por dw = 2(1- ρ). Assim, ρ = (2-dw)/2 e, em ausência de autocorrelação espacial, dw tem valor esperado igual a 2. Valores de dw superiores a 2 indicam autocorrelação negativa ao passo que valores de dw inferiores a 2 indicam autocorrelação positiva. Em geral, a heterogeneidade espacial causa autocorrelações positivas, ao passo que os efeitos de competição entre plantas vizinhas causam autocorrelações negativas.

4. Métodos de Análise Espacial e Áreas de Aplicação em Genética

Os principais métodos de análise espacial são:

(a) método das variáveis regionalizadas ou procedimentos geoestatísticos;

(b) métodos de análise de séries temporais;

(c) modelo linear de campo aleatório;

(d) análise de superfícies de resposta polinomiais.

O método de análise de superfícies de tendências baseia-se no ajuste de superfícies de resposta polinomiais ou de "splines" e foi recomendado por Federer (1998). Entretanto, segundo Gilmour (2000), tais modelos raramente se justificam na prática. Verbyla et al. (1999) reportam que a tendência global pode ser acomodada através do ajuste de "splines" cúbicas. O ajuste de tais "splines" pode ser realizado através do software ASREML (Gilmour et al., 2000).

O modelo linear de campo aleatório foi desenvolvido por Zimmermann & Harville (1991) e considera diretamente a heterogeneidade espacial através da inclusão dos efeitos de tendência e das correlações entre os erros, modelando de forma direta a matriz de covariância dos erros. Esta modelagem é similar à empregada nos métodos geoestatísticos para fazer predições espaciais pois consideram as observações como sendo uma realização parcial de um campo aleatório. O método considera dependência espacial em todas as direções e preserva quaisquer esquemas de blocagem. Tal modelo procura uma estimativa da função geral de covariância, a qual participa diretamente dos procedimentos de estimação e predição.

Esta função é da forma $R = \sigma^2 \left[f(h) \right]$, em que f (h) é uma função da distância entre as observações nas posições físicas s e (s + h), em que cada posição é definida por um par de coordenadas dado pela linha e coluna na grade experimental. Tem-se $Cov\ (e_s,\ e_{(s+h)}) = Cov\ (h) = \sigma^2$ se $h = 0$ e $Cov\ (e_s,\ e_{(s+h)}) = \sigma^2 \left[f(h) \right]$, se h > 0. A determinação da função geral que descreve a covariância espacial pode ser realizada pela abordagem geoestatística, ajustando uma função contínua ao variograma amostral, conforme detalhado na seqüência.

No método das variáveis regionalizadas utilizam-se os procedimentos geoestatísticos para avaliar o padrão de variabilidade espacial na área experimental obtendo-se a matriz de semivariâncias ajustadas, a qual é utilizada como ponderadora no sistema de equações de quadrados mínimos generalizados (Cressie, 1993; Martinez, 1994) ou modelo misto. Martinez (1994) empregou tal método usando semivariâncias ajustadas pelo modelo esférico.

As principais ferramentas geoestatísticas utilizadas na análise espacial são (Valente, 1989; Ribeiro Júnior, 1995; Landim, 1998; Duarte, 2000; Soares, 2000; Vieira, 2000):

(i) Cômputo das semivariâncias

A semivariância é definida como $S(h) = (1/2)\ Var\left[e_{(s+h)} - e_{(s)}\right]$, ou seja, eqüivale à metade da variância de diferenças entre observações (entre erros no caso) separadas por uma distância h. Valores baixos de S(h) indicam menor variabilidade, ou seja, maior similaridade.

Dentre os vários estimadores de semivariâncias, o mais utilizado é o estimador clássico de Matheron, baseado no método dos momentos e dado por:

$\hat{S}(h) = \frac{1}{2N(h)} \sum_{N(h)} \left[\hat{e}_{(s+h)} - \hat{e}_{(s)}\right]^2$, em que N(h) é o número de diferenças tomadas à distância h.

(ii) Construção de semivariogramas ou variogramas

O semivariograma ou variograma representa uma função de semivariâncias em relação às suas respectivas distâncias e é utilizado para a estimação da estrutura de variabilidade espacial como função da distância entre as observações. O variograma amostral é construído plotando-se os valores calculados de semivariância no eixo das ordenadas contra as respectivas distâncias no eixo das abcissas, ou seja, plotando-se os pontos $\left[h, \hat{S}(h)\right]$.

Se os valores de $\hat{S}(h)$ distribuem-se aleatoriamente em função de h pode-se inferir sobre independência entre observações ou resíduos. Por outro lado, se $\hat{S}(h)$ crescem com o aumento de h pode-se inferir sobre dependência espacial entre observações ou resíduos, revelada pela menor variabilidade à menores distâncias.

(iii) Realização de Krigagem

O semivariograma amostral não é suficiente para descrever a variabilidade em toda a área, uma vez que baseia-se em estimativas de semivariâncias apenas para algumas distâncias. Uma interpolação por Krigagem permite descrever a variabilidade para toda a área, tendo por base a estimativa da estrutura de

dependência entre observações vizinhas e o ajuste de uma função contínua de semivariograma a partir do semivariograma amostral discreto.

(iv) Obtenção da função que descreve a covariância espacial

A partir do ajuste de uma função contínua ao variograma amostral obtém-se a função de covariância espacial $Cov\ (e_s,\ e_{(s+h)})$ pela relação $Cov\ (e_s,\ e_{(s+h)}) = Cov\ (h) = \sigma^2 - S(h)$, em que $\sigma^2 = Cov\ (h=0)$. As funções de variograma mais comumente utilizadas são os modelos esférico, exponencial e gaussiano, os quais correspondem às seguintes funções de autocovariância:.

Modelo esférico

$$Cov\ (e_s,\ e_{(s+h)}) = \begin{cases} \sigma^2 \left[1 - \frac{3}{2}\left(\frac{h}{\alpha}\right) + \frac{1}{2}\left(\frac{h}{\alpha}\right)^3\right] \\ 0,\ se\ h \geq \alpha \end{cases}, \text{ se } h > a$$

Modelo exponencial

$$Cov\ (e_s,\ e_{(s+h)}) = \sigma^2\ \exp\left(\frac{-h}{\alpha_e}\right),\ em\ que\ \alpha_e = \frac{1}{3}\ \alpha$$

Modelo gaussiano

$$Cov\ (e_s,\ e_{(s+h)}) = \sigma^2\ \exp\left(\frac{-h^2}{\alpha^2}\right),$$ em que α refere-se ao alcance da correlação espacial ou à distância a partir da qual a semivariância estabiliza.

Os métodos de análise de séries temporais foram utilizados inicialmente por Gleeson e Cullis (1987) que consideraram os erros através de um processo autoregressivo integrado de médias móveis (ARIMA (p, q, d) em uma única direção. Este modelo foi considerado ineficiente por considerar os efeitos espaciais em uma só dimensão e Martin (1990) e Cullis & Gleeson (1991) estenderam tal modelo a duas dimensões: linhas e colunas. Tal modelo estendido é da forma ARIMA (p_1, d_1, q_1) x ARIMA (p_2, d_2, q_2).

Estes modelos são denominados modelos com erros nas variáveis e consideram um efeito de tendência *(*ξ*)* mais um erro η independente. A diferença entre os vários métodos reside na modelagem e estimação do efeito de tendência. Nos métodos mencionados, a tendência é modelada por um processo ARIMA, o qual limita-se a diferenças de primeira ordem ou diferenças entre observações adjacentes. Alguns métodos somente incluem ξ e ignoram η.

Nos ensaios com culturas agrícolas, na ausência de conhecimento da correta estrutura de correlação, Gilmour et al. (1997) sugerem a modelagem de ξ como um processo auto-regressivo separável de primeira ordem (AR1 x AR1), ignorando η. Este processo auto-regressivo bidimensional de primeira ordem tem-se mostrado eficiente em uma gama de situações (Apiolaza et al., 2000; Gilmour, 2000; Costa e Silva et al., 2001), algumas com a inclusão de η.

Os métodos de análise espacial têm, recentemente, sido aplicados a várias sub-áreas da genética e melhoramento. No Brasil, trabalho relevante foi realizado por Duarte (2000), aplicando e comprovando a eficiência da análise espacial via métodos geoestatísticos no contexto de melhoramento de plantas anuais, através de uma implementação via o Programa SAS (Statistical Analysis System). Na Austrália, Gilmour & Thompson (1998) e Gilmour et al. (2000) implementaram o software ASREML, destinado a análise espacial de experimentos em melhoramento de plantas anuais e perenes, permitindo a análise ao nível de indivíduos e o emprego de vários procedimentos de análise espacial com ênfase nos métodos de análise de séries temporais.

No contexto da análise de medidas repetidas com dependência temporal, o relevante trabalho de Apiolaza et al. (2000) demonstrou a utilidade dos modelos auto-regressivos com variâncias heterogêneas. Os procedimentos de análise espacial têm sido aplicados também na análise de dados moleculares, associados à genética de populações separadas espacialmente (Degen &

Scholz, 1998; Lecorre et al., 1998) e também na análise de QTL´s considerando os efeitos de posição no cromossomo, revelando a grande utilidade da abordagem espacial para a genética e melhoramento.

Tendo em vista os bons resultados práticos propiciados pelos procedimentos geoestatísticos e de análise de séries temporais, este artigo enfatiza mais estes dois enfoques. Ênfase também é dada (Tópico 8) à análise espacial ao nível de indivíduos via abordagens de modelos lineares mistos (REML/BLUP) visando a estimação de componentes de variância e a predição de valores genéticos no melhoramento de plantas perenes.

5. Noções de Processos Estocásticos

As técnicas de análise de séries temporais e de análise geoestatística dependem fundamentalmente da disciplina de processos estocásticos. Assim, algumas noções de processos estocásticos são apresentadas neste tópico, com base em Papoulis (1981).

5.1. Processo estocástico e variável aleatória

Um processo estocástico $Z(w,t)$ é uma função de duas variáveis w e t, cujos domínios são Ω e $\Re$ ($w \in \Omega$, $t \in \Re$). Se $t \in I \subset \Re$ e se I contém um número contável ou enumerável de elementos, $Z(w,t)$ é discreto. Se I é infinito e não enumerável, $Z(w,t)$ é contínuo. Fixando a variável t, w é uma variável aleatória. (uma variável aleatória é função de apenas uma variável).

Um processo estocástico real é Gaussiano se as variáveis aleatórias $Z(t_1)$, $Z(t_2)$, ..., $Z(t_n)$ são conjuntamente normais para qualquer n e t_1, t_2, ... t_n. A identificação de um processo estocástico se dá via a sua função de autocorrelação e/ou função de autocorrelação parcial.

5.2 Função de autocorrelação de um processo estocástico ($\rho_Z(t_1, t_2)$)

A autocorrelação é a correlação de Z com ele mesmo mas em distintos tempos ou espaços. A autocorrelação de um processo estocástico é dada por:

$$\rho_Z(t_1, t_2) = E[Z(t_1) \cdot Z(t_2)] \quad \forall (t_1, t_2)$$

5.3. Função de autocovariância de um processo estocástico ($\Upsilon_Z(t_1, t_2)$) *e relação com* ($\rho_Z(t_1, t_2)$)

A função de autocovariância de um processo estocástico é dada por:

$$\Upsilon_Z(t_1, t_2) = E[Z(t_1) - E(Z(t_1))]\ [Z(t_2)-E(Z(t_2))] \quad \forall (t_1, t_2)$$

As relações entre $\rho_Z(t_1, t_2)$ e $\Upsilon_Z(t_1, t_2)$ são:

(i) $\Upsilon_Z(t_1, t_2) = \rho_Z(t_1, t_2) - E(Z(t_1)) \cdot E(Z(t_2))$

(ii) $Var(Z_t) = \rho_Z(t_1, t_2) - [E(Z(t))]^2$, em que $E(Z(t))$ é a esperança ou média do processo estocástico.

5.4. Estacionariedade dos processos estocásticos

Seja o processo estocástico $Z(t)$ tal que, fixando o valor de t, tem-se que $Z(t)$ é uma variável aleatória Z_t cuja função densidade de probabilidade é:

$$f_{Z_t}(z) = \frac{1}{\sqrt{2\pi(t^2+t+1)}}\ e^{-\frac{1}{2}\frac{z^2}{(t^2+t+1)}} \qquad z \in \Re$$

Observa-se que $f_{Zt}(z)$ depende de t. Assim, $Z(t)$ é um processo estocástico não estacionário de primeira ordem. Caso a função densidade de probabilidade não dependesse de t, $Z(t)$ seria um processo estocástico estacionário de primeira ordem.

Para se fazer predições é necessário a condição de estacionariedade. A estacionariedade de um processo estocástico pode ser no sentido restrito ou amplo.

Seja um processo estocástico $Z(t)$ e considere os n instantes $t_1, t_2, \ldots t_n$. Se para qualquer número $n \in I$ de variáveis aleatórias $Z_{t_1}, Z_{t_2}, \ldots Z_{t_n}$ a função densidade de probabilidade conjunta de ordem n

$$f_{Z_{t1},\ Z_{t2},\ ...,\ Z_{tn}}(z_1,\ z_2,\ ...,\ z_n) = f_{Z_{t1+\tau},\ Z_{t2+\tau},\ ...\ Z_{tn+\tau}}(z_1,\ z_2,\ ...,\ z_n)$$

não variar com um deslocamento no tempo ou espaço, então o processo estocástico é dito estacionário no sentido restrito ($Z(t)_r$). Neste caso, a distribuição é a mesma em qualquer tempo.

Por outro lado, um processo estocástico é estacionário no sentido amplo ($Z(t)_a$) se $E[Z(t_1)] = \mu(t) = \eta \ \forall t$ e $\rho_Z(t_1, t_2) = \rho_Z(\tau)$ com $\tau = (t_2 - t_1)$. Neste caso, a média é constante (não depende de t) e a função de autocorrelação depende só de $t_2 - t_1$ (intervalo de tempo). Em resumo, em um processo estocástico $Z(t)_r$ a função é constante e, em um processo $Z(t)_a$ a média é constante e a função de autocorrelação só depende da diferença entre tempos.

5.5 Ergodicidade dos processos estocásticos

A caracterização completa de um processo estocástico exige o conhecimento de todas as suas funções amostras ou realizações do processo. Isto permite determinar a média e a função de autocorrelação. Para alguns processos estocásticos, denominados ergódicos, estes parâmetros podem ser determinados a partir de apenas uma função amostra típica do processo. Para os processos ergódicos, os valores médios e momentos podem também ser determinados através de médias temporais. Um caso típico de processo estocástico ergódico refere-se às séries temporais.

Uma série temporal não pode ser estudada via regressão, uma vez que esta técnica supõe independência entre $t_1, t_2, \ldots, t_n$, o que não é realístico para tal série. Neste contexto, lança-se mão das técnicas de análise de séries temporais que, geralmente, assumem estacionariedade. Entretanto, quando não se tem estacionariedade, trabalha-se com diferenças entre os tempos, podendo-se diferenciar os dados várias vezes.

Dado o processo estocástico real $Z(t)$ e querendo-se estimar a sua média $\eta = E(Z(t))$, pode-se formar a média temporal:

$\eta_T = \frac{1}{2T}\int_{-T}^{T} Z(t)\ dt$, em que η_T é uma variável aleatória com média

$$E(\eta_T) = \frac{1}{2T}\int_{-T}^{T} E[Z(t)]\ dt = \eta$$

Para estabelecer a ergodicidade de um processo é suficiente *achar Var (η_T)* e examinar as condições sob os quais *Var (T)* $\rightarrow 0$, quando $T \rightarrow \infty$. Se Var *(η_T)* $\rightarrow 0$, quando $t \rightarrow \infty$, então $\eta_T = \eta$ e o processo estocástico é ergódico na média. Dessa forma, *Z(t)* é ergódico na média se $\eta_T \rightarrow \eta$ quando $T \rightarrow \infty$ e a ergodicidade da média é a versão no tempo da lei dos grandes números (lei fraca: convergência em distribuição; lei forte: convergência em probabilidade).

6. Métodos Geoestatísticos

6.1 Origem e fundamentos da geoestatística

Na África do Sul, o engenheiro de minas D.G. Krige e o estatístico H.S. Sichel desenvolveram uma técnica de estimação, a qual, posteriormente, recebeu um tratamento formal por Matheron no Centre de Morphologie Mathematique, em Fountainebleau, França. Tal metodologia recebeu o nome de Geoestatística, disciplina esta que procura estudar o comportamento das chamadas **variáveis regionalizadas**, ou seja, variáveis com comportamento espacial, as quais mostram características intermediárias entre as variáveis verdadeiramente casuais ou aleatórias e aquelas totalmente determinísticas. Os trabalhos iniciais de Mathernon datam de 1962 e 1963, sendo que o trabalho clássico deste autor foi traduzido para o inglês em 1971. No Brasil, as primeiras publicações datam dos anos 80, citando-se por exemplo, Valente (1989).

As variáveis regionalizadas apresentam uma aparente continuidade no espaço. A continuidade geográfica se manifesta pela tendência que a variável tem de apresentar valores muito próximos (dependentes) em dois pontos vizinhos e diferentes em pontos distantes. O grau de dependência em relação as posições vizinhas é usualmente expresso em termos da memória envolvida no processo. Nos modelos determinísticos de dependência, a observação na posição s depende de todas as posições prévias ou próximas, ou seja, o processo é de longa memória. Por outro lado, os processos puramente aleatórios não apresentam memória, sendo independentes da seqüência de posições. Quando

se utiliza a estatística clássica para representar as propriedades dos valores amostrais, assume-se que estes são realizações de uma variável casual e as posições relativas das amostras são ignoradas.

As variáveis regionalizadas não são realizações de uma variável aleatória, pois são correlacionadas. Por outro lado, não atendem completamente os requisitos de um processo estocástico ergódico e estacionário, pois constituem uma só realização. O conceito de estacionariedade do modelo das funções aleatórias, apesar de teoricamente imprescindível para qualquer ação de inferência estatística, não é validável ou refutável a priori, visto que se conhece uma só realização da função, dada pelo conjunto de dados espacialmente distribuídos. Entretanto, a geoestatística não toma esta limitação como restritiva, mas parte do princípio de que a hipótese de estacionariedade pode ser julgada como apropriada para o conjunto de dados. Dessa forma, a hipótese de estacionariedade da média é parte integrante e fundamental do modelo probabilista geoestatístico (Soares, 2000). Esta restrição estacionária é semelhante, em termos de concepção, à ergodicidade nas séries temporais dependentes. Esta restrição é denominada **hipótese intrínseca**. Uma variável aleatória estacionária é sempre ergódica com respeito à sua média e à sua função de autocovariância.

Concebidos juntos com a própria geoestatística surgiram os conceitos de **variograma** como medidor da continuidade espacial, de **anisotropia** espacial quando a continuidade varia nas diferentes direções ou dimensões do espaço, de **efeito pepita (nugget)** ou variabilidade à pequena escala ou localizada, dentre outros.

Na seqüência são descritos alguns conceitos e métodos em geoestatística com base em Landim (1998) e Soares (2000).

6.2. Semivariância e autocovariância espacial

Uma observação localizada espacialmente em posição x_1, denominação genérica de um conjunto de coordenadas geográficas, é interpretado como uma realização $z(x_1)$ de uma variável aleatória $Z(x_1)$. No espaço A, no qual se dispensa o conjunto de pontos amostrados, tem-se as realizações das N variáveis aleatórias $Z(x_1)$, $Z(x_2)$, ¼ $Z(x_N)$, correlacionadas entre si.

Para cada uma das variáveis aleatórias, define-se o primeiro (média) e o segundo (variância) momentos:

$$E[Z(x_i)] = m(x_i) = \int_{-\infty}^{+\infty} z dF_{x_i}(j) = \int_{-\infty}^{+\infty} z f_{x_i}(z) dz$$

$$Var[Z(x_i)] = \int_{-\infty}^{+\infty} [z - m(x_i)]^2 dF_{x_i}(z)$$, em que:

$f_{x_i}(z)$: função densidade de probabilidade da variável aleatória *Z(x$_i$);*

$F_{x_i}(z)$: função distribuição acumulada da variável aleatória *Z(x$_i$).*

A covariância entre duas variáveis *Z(x$_1$)* e *Z(x$_2$)* é definida como:

Cov[*Z(x$_1$), Z(x$_2$)*] = *E*[*Z(x$_1$), Z(x$_2$)*] - *m(x$_1$) m(x$_2$),* em que:

$$E[Z(x_1)Z(x_2)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy d^2 F_{\chi_i,\chi_2}(x,y)$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_{x_i,x_2}(x,y)\ dx\ dy;$$

$F_{x1,\ x2}\ (x,y)$*:* função distribuição bivariada dada por:

$$F_{x_1,x_2}(x,y) = prob\ [Z(x_1) \le x\ e\ Z(x_2) \le y].$$

O coeficiente de correlação entre as variáveis *Z(x$_1$)* e *Z(x$_2$)* é dado por:

$$\rho[Z(x_1),\ Z(x_2)] = \frac{Cov[Z(x_1),\ Z(x_2)]}{[Var[Z(x_1)]Var[Z(x_2)]]^{1/2}}$$

A semivariância é definida como:

$$\lambda\left[Z(x_1),\ Z(x_2)\right] = E\left\{\left[Z(x_1) - Z(x_2)\right]^2\right\}.$$

O conjunto de variáveis aleatórias $Z(x_i)$ correlacionadas entre si, constituem uma função aleatória da qual só se conhece uma realização. A hipótese da estacionariedade em relação aos parâmetros descritos anteriormente é definida pela seguinte afirmativa: a correlação entre duas variáveis aleatórias depende somente da distância espacial que as separa e é independente de sua localização. Assim, para duas variáveis aleatórias distanciadas espacialmente por um vetor h, com direção e orientação específica em um espaço de uma, duas ou três dimensões, tem-se que a semivariância $(\lambda(h))$ e a autocovariância espacial $(\gamma(h))$ dependem de h, ou seja:

$$Cov[Z(x_1), Z(x_2)] = Cov[Z(x_i), Z(x_i + h)] = \gamma(h)$$
$$\lambda[Z(x_1), Z(x_2)] = \lambda[Z(x_i), Z(x_i + h)] = \lambda(h)$$

Tal como para a média, que pode ser estimada por uma média aritmética dos valores das realizações das variáveis aleatórias, ou seja, por

$m = \frac{1}{N}\sum_{i=1}^{N} Z(x_i)$, supondo estacionariedade da média ou do primeiro

momento, a hipótese de estacionariedade do segundo momento permite inferências sobre a semivariância e sobre a autocovariância espacial via:

$$\gamma(h) = \frac{1}{N(h)}\sum_{i=1}^{N(h)} \left[Z(x_i).Z(x_i+h)\right] - m\ (x_i)\ m\ (x_i+h)$$

$$\lambda(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)} \left[Z(x_i) - Z(x_i+h)\right]^2$$, em que:

N(h): número de pontos ou diferenças $\left[Z(x_i+h) - Z(x_i)\right]$ tomados à distância *h*, com $Z(x_i)$ referindo-se a observação realizada na posição x_i e $Z(x_i + h)$ a observação realizada na posição $x_i + h$.

A hipótese intrínseca estabelece que $\left[(x_i)-(x_i+h)\right]$ é estacionária de segunda ordem e λ *(h)* representa a semivariância ou função intrínseca. Por outro lado, pela estacionariedade de primeira ordem ou do primeiro momento, $E\left[(x_i)-(x_i+h)\right]=m(h)$ *e* $m(h)$ representa a tendência.

No caso de variáveis regionalizadas com estacionariedade de segunda ordem, tem-se a relação:

$\sigma^2_{Z_{(x_i)}} = \lambda(h)+\gamma(h)$, em que:

$\sigma^2_{Z_{(x_i)}}$: variância populacional;

$\lambda(h)$: semivariância para uma distância *h*;

$\gamma(h)$: autocovariância para uma distância *h*;

$\sigma^2_{Z_{(x_i)}} = \lambda(h)+\gamma(h)$: variância populacional = semivariância + autocovariância para uma distância h.

De modo geral, estes parâmetros e hipóteses de estacionariedade são os constituintes fundamentais dos modelos geoestatísticos. Os valores das amostras são realizações de variáveis aleatórias localizadas espacialmente em *A* e a autocovariância e a semivariância, por não dependerem da localização das variáveis mas somente do vetor *h*, medem a continuidade espacial ou dispersão das variáveis no espaço.

6.3 Semivariograma e correlograma

Para uma variável quantitativa *Z(x)*, os diagramas de representação dos pares de pontos amostrais *Z(x)* nos eixos *x* e *Z(x + h)* no eixo *y*, para diferentes valores de *h*, contêm informações muito ricas sobre a continuidade espacial de uma área. Neste caso, em geral, para pequenos valores de *h*, observa-se pela distribuição dos pontos ao redor de uma reta de 45º com eixo de *x,* uma boa correlação linear espacial entre amostras distanciadas a pequenos intervalos em determinada direção. À medida que os valores de *h* aumentam, as nuvens de pontos tornam-se mais dispersas, revelando uma diminuição da correlação espacial entre amostras. No caso de experimentos na área agrícola e florestal é recomendável a

representação dos erros estimados associados a vários pontos amostrados, ou seja, a obtenção de diagramas *e(x)* versus *e(x + h)*.

O conjunto de vários diagramas ou gráficos, para os diferentes valores de *h*, contêm a quase totalidade da informação disponível sobre o grau de dispersão/ continuidade da variável *Z(x).* Entretanto, para uma melhor interpretação e síntese da informação em função da evolução de h pode-se lançar mão do semivariograma, o qual refere-se a disposição gráfica dos valores de h no eixo *x* e dos valores de semivariâncias *(λ(h))* no eixo *y*. Podem ser construídos semivariogramas considerando várias direções de *h* no espaço, sendo possível que *Z(x)* não apresente o mesmo comportamento em todas as direções, em termos de correlação espacial.

Uma outra medida de continuidade espacial é fornecida pela média dos produtos *z(x) z(x + h).* Esta média dos produtos pode ser um estimador da covariância não centrada, equivalente a: $Cov'(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(x_i)\ Z(x_i+h)]$ ou um estimador da covariância centrada ou autocovariância $\gamma(h)$ ou ainda um estimador da autocorrelação que é a forma normalizada da covariância. O estimador da autocorrelação no intervalo h é dado por:

$$\rho(h) = \frac{\gamma(h)}{\left[Var[(x_i)\ Var(x_i+h)\right]^{1/2}}$$, em que:

$$Var(x_i) = \sigma^2_{(x_i)} = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - n\ (x_i)]^2 ;$$

$$Var(x_i+h) = \sigma^2_{(x_i+h)} = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(x_i+h) - m\ (x_i+h)]^2 .$$

O **correlograma** refere-se a representação gráfica dos pontos [*(h),* ρ*(h)*] no plano cartesiano *x, y*. O semivariograma relaciona-se com o correlograma da seguinte maneira: $\lambda(h) = \sigma^2_{x_i}[1-\rho)(h)]$, ou seja, $\rho(h) = 1 - \lambda(h) / \sigma^2_{x_i}$.

Conforme Ribeiro Júnior (1995), os semivariogramas exigem hipóteses de estacionariedade menos restritivas, abrangendo um maior número de situações, sendo por isso preferidos em relação aos correlogramas e aos covariogramas. O

coeficiente de autocorrelação depende da variância, que não pode ser infinita e, como a semivariância é livre desta restrição, torna-se preferida. Além do mais, conforme Landim (1998), o uso da semivariância revela com mais facilidade a presença de tendência nos dados.

Assumindo as hipóteses de estacionariedade dos acréscimos h, tem-se:

$\lambda(h) = \frac{1}{2} \; E\;[Z(x) - Z(x+h)]^2$: semivariância ou metade da variância de diferença entre observações numa variável aleatória Z, separadas por uma distância h;

$\gamma(h) = E\;[Z(x)\;\; Z(x+h)] - E\;[Z(x)]\;E\;[Z(x+h)]$: autocovariância centrada.

Para funções aleatórias estacionárias e admitindo estacionariedade do primeiro momento, tem-se:

$$\gamma(h) = E\;[Z(x).Z(x+h)] - m^2$$

$$\lambda(h) = E\;[Z(x)]^2 - E\;[Z(x)\;\; Z(x+h)]$$

Assim, deduz-se a seguinte relação entre semivariância e autocovariância:

$\lambda(h) = \gamma(0) - \gamma(h)$ ou

$\gamma(0) = \sigma^2 = \lambda(h) + \gamma(h)$

A autocorrelação é então dada por:

$$\rho(h) = \gamma(h)/\gamma(0) = \gamma(h)/Var[Z(x_i)] = \gamma(h)/\sigma^2_{Z(x_i)}$$

Para malhas regulares, como a dos experimentos de campo, o semivariograma experimental é obtido conforme os seguintes passos (Ribeiro Júnior, 1995; Duarte, 2000):

(i) fixa-se uma distância h ou "lag";

(ii) formam-se todos os pontos separados pela distância h;

(iii) aplica-se a expressão do estimador para se obter a semivariância associada à distância h;

(iv) toma-se outra distância ou "lag" e repetem-se os passos de (i) a (iii), o que deve ser feito até uma distância máxima de interesse;

(v) obtém-se o semivariograma, plotando-se os pontos formados pelas distâncias no eixo x e pelas semivariâncias estimadas, no eixo y.

6.4. Modelos para os semivariogramas ou variogramas

Conforme Landim (1998), a semivariância não é apenas igual à média das diferenças ao quadrado entre pares de pontos espaçados pela distância h, mas também é igual a variância destas diferenças. Assim, os semivariogramas são denominados também de variogramas.

A semivariância $\lambda(h)$ pode ser avaliada somente a distância h correspondentes a múltiplos do espaçamento entre pontos de amostragem ao longo da direção considerada. O vetor h apresentando-se infinitamente pequeno faz com que a variância e a covariância ou autocovariância se tornem muito próximas. Para valores grandes de h, a covariância diminuirá ao passo que a variância aumentará. Dessa forma, a semivariância distribui-se de 0, quando $h = 0$, até um valor igual a variância das observações, para um valor alto de h. A distância na qual $\lambda(h)$ atinge um patamar igual a variância dos dados, patamar este denominado **soleira** (sill), é chamado **alcance**. A soleira é simbolizada por *C* e o alcance por α.

A variável regionalizada é composta de duas partes: a tendência e o resíduo. A tendência é o valor esperado da variável regionalizada em um determinado ponto x_i, que eqüivale à média ponderada de todos os pontos em torno de uma vizinhança x_i. Subtraindo a tendência, da variável regionalizada, os próprios resíduos serão a variável regionalizada estacionária. A construção do

semivariograma pode ser baseada nos dados reais ou nos resíduos e faz parte da análise estrutural em geoestatística.

Os semivariogramas expressam o comportamento espacial da variável regionalizada e informam sobre:

(i) padrão de variabilidade nas várias direções: tem-se uma **isotropia** quando o padrão de variabilidade é o mesmo em todas as direções, gerando semivariograma omnidirecional; tem-se anisotropia quando o padrão de variabilidade difere em função das direções, requerendo semivariogramas direcionais;

(ii) efeito pepita ou nugget: ocorre quando para h = 0, a semivariância $\lambda(h)$ já apresenta algum valor, quando deveria ser nula revelando similaridade absoluta à distância nula.

O efeito pepita é simbolizado por C_o e pode ser atribuído a erros de medição ou ao fato dos dados não terem sido coletados a intervalos suficientemente pequenos para exibir o comportamento espacial do fenômeno estudado. O efeito pepita significativo denota que há uma grande variabilidade à pequena escala.

O efeito C_o mede fundamentalmente duas parcelas da variabilidade total do fenômeno: (a) a variabilidade correspondente a uma pequena escala não abrangida pela malha de amostragem; (b) a variabilidade induzida por erros não sistemáticos de amostragem, os quais acrescentam um ruído branco ou aleatório.

(iii) forma da variabilidade espacial: em manchas, em gradientes ou completamente aleatória.

O gráfico formado por $[(h), \lambda(h)]$ é denominado semivariograma experimental. O grau de aleatoriedade presente nos dados é dado pela expressão $r = C_o/C$ e pode ser interpretado da seguinte maneira:

r < 0,15: componente aleatória pequena;
0,15 < r < 0,30: componente aleatória significativa;
r > 0,30: componente aleatória muito significativa.

Após a obtenção do semivariograma experimental é necessário ajustá-lo a um modelo teórico. Ajustar um semivariograma através de uma curva média permite inferir sobre o comportamento de $\lambda(h)$ representativo para toda a área e gama de valores de h. Dentre os valores teóricos de semivariogramas, os principais são:

(1). Modelos com soleira

(1.a) Modelo esférico

Trata-se de um dos modelos mais usuais em geoestatística, o qual é função de dois parâmetros: o patamar ou soleira *C* e a amplitude ou alcance $h = \alpha$. Tal modelo tem a seguinte expressão:

$$\lambda(h) = \begin{cases} C\left[1{,}5\ \frac{h}{a} - 0{,}5\left(\frac{h}{\alpha}\right)^3\right] & para\ h \leq \alpha \\ C & para\ h > \alpha \end{cases}$$

É um modelo padrão assim como é a distribuição normal para a estatística clássica.

(1.b) Modelo exponencial

É função também do patamar e da amplitude, sendo dado por:

$$\lambda(h) = C\ \left[1 - e^{-3h/\alpha}\right]$$

(1.c) Modelo gaussiano

Este modelo tem a seguinte expressão:

$$\lambda(h) = C\ \left(1 - \exp\left(\frac{-3h^2}{\alpha^2}\right)\right)$$

(2) Modelos sem soleira

(2.a) Modelo linear

É o modelo mais simples, dado por: $\lambda(h) = ph$ em que p é a inclinação da reta. Uma extensão deste modelo é $\lambda(h) = ph^{\alpha}$.

(2.b) Modelo Wysianiano

Neste modelo, dado por: $\lambda(h) = 3\alpha \ \log_e(h)$, o semivariograma torna-se linear se utilizar o logaritmo da distância h.

6.5. Krigagem e Cokrigagem

A Krigagem é um processo de estimação de valores de variáveis distribuídas no espaço, a partir de valores adjascentes, considerados dependentes pelo semivariograma. Trata-se, em essência, de um método de estimação por médias móveis. Esta técnica presta-se, dentre outras utilidades, para a estimação da tendência, de modo similar à técnica de análises de superfícies de tendência.

Além da estimativa, a Krigagem fornece também o erro de estimação. A técnica usa informações do semivariograma para encontrar os pesos ótimos das amostras que irão estimar um ponto, sendo que os pesos serão diferentes conforme o seu arranjo geográfico. As formas mais usuais do método são denominadas Krigagem simples, Krigagem ordinária, Krigagem universal e Krigagem intrínseca.

A Krigagem ordinária é uma estimação linear para uma variável regionalizada que satisfaz a hipótese intrínseca, mas que não exige que a média seja conhecida. Por outro lado, a Krigagem simples, sob a hipótese da estacionariedade exige que a média seja conhecida. Assim, a Krigagem simples tem uma relação direta com a estimação por máxima verossimilhança (ML) e a Krigagem ordinária tem uma relação direta com a estimação por verossimilhança restrita (REML). A Krigagem ordinária, também denominada Krigagem normal, é um estimador linear geoestatístico não viesado e com mínima variância de estimação, ou seja, é um estimador BLUE.

A Krigagem ordinária é usada quando a variável regionalizada é estacionária de primeira ordem. Para variáveis não estacionárias, ou seja, com tendência, mas para cujos erros a hipótese intrínseca se verifica, recomenda-se a utilização da Krigagem universal.

A Cokrigagem é um método geoestatístico segundo o qual diversas variáveis regionalizadas podem ser estimadas simultaneamente, tendo por base a correlação espacial entre elas. Refere-se, portanto, a uma extensão multivariada da Krigagem, quando se trabalha com um vetor de valores para cada ponto amostrado e não com um único valor.

7. Análise de Séries Temporais

7.1. Conceitos

Uma série temporal refere-se a um conjunto de observações ordenadas segundo a variável tempo e eqüivale à própria função amostra ou realização de um processo estocástico. A série é, então, uma realização dentre muitas que poderiam ter sido observadas e a variável *t* em *Z(t),* referida como tempo, pode também ser função de um parâmetro físico como o espaço.

A decomposição clássica de uma série temporal é dada por:

$Z(t) = T(t) + C(t) + S(t) + a(t)$, em que:

T(t): tendência em função do tempo;
C(t): ciclo em função do tempo;
S(t): sazonalidade em função do tempo;
a(t): ruído branco aleatório.

Um dos objetivos da análise de séries temporais é a verificação de tendências, ciclos e variações sazonais. Assim, modelos probabilísticos ou estocásticos são construídos no domínio temporal e estes devem ser simples e parcimoniosos no sentido que o número de parâmetros envolvidos deve ser o menor possível. Em geral, nos ensaios agrícolas considera-se o modelo $Z(t) = T(t) + a(t)$, que é função apenas da tendência e erro aleatório.

Freqüentemente, supõe-se que uma série temporal é estacionária, ou seja, que ela se desenvolve no tempo, aleatoriamente ao redor de uma média constante,

refletindo alguma forma de equilíbrio estável. Entretanto, algumas séries encontradas na prática apresentam alguma forma de não estacionariedade, que pode ser explosiva ou não explosiva. Uma classe de modelos paramétricos denominada ARIMA é capaz de descrever satisfatoriamente séries estacionárias e não estacionárias do tipo não explosiva ou não intensiva (não estacionariedade homogênea) (Morettin e Toloi, 1987).

Para séries não estacionárias é necessário transformar os dados originais visando a obtenção de estacionariedade. A transformação mais comum consiste em tomar diferenças sucessivas da série original. A primeira diferença de $Z(t)$ é definida por $\Delta Z(t) = Z(t) - Z(t-1)$ e a segunda diferença é dada por $\Delta^2 Z(t) = \Delta[\Delta Z(t)] = \Delta[Z(t)-Z(t-1)] = Z(t) - 2Z(t-1) + Z(t-2)$. Em geral, a n-ésima diferença de $Z(t)$ é $\Delta^n Z(t) = \Delta[\Delta^{n-1} Z(t)]$]. Em situações normais, tomando-se uma ou duas diferenças, a série se torna estacionária (Morettin e Toloi, 1987).

7.2. Modelos para a tendência

Na análise de séries temporais, a escolha da estrutura do modelo baseia-se nos próprios dados e é baseada nos seguintes estágios de um ciclo iterativo (Morettin e Toloi, 1987):

(i) especificação: uma classe geral de modelos é considerada para análise;

(ii) identificação: baseada na análise de autocorrelações e autocorrelações parciais (é o estágio mais crítico);

(iii) estimação: parâmetros do modelo identificado são estimados;

(iv) verificação: através de análise dos resíduos verifica-se se o modelo ajustado presta-se aos fins de previsão ou predição.

Dentre os modelos paramétricos de análise de séries temporais, o método de Box e Jenkins tem sido um dos mais usados (Box e Jenkins, 1970). Este método consiste em ajustar modelos auto-regressivos-integrados-médias móveis, ARIMA (p, d, q) a um conjunto de dados. No caso, p, d e q referem-se a ordem do processo auto-regressivo, número de diferenças ou de diferenciações para tornar a série estacionária e ordem do processo médias móveis, respectivamente. Dependendo dos valores de p, q e d, os modelos ARIMA são equivalentes aos modelos: médias móveis (MA), quando $p = d = 0$; auto-regressivos (AR)

quando d = q = 0; auto-regressivo-médias móveis (ARMA) quando d = 0. Assim, tem-se:

ARIMA (p, 0, 0) = AR(p);

ARIMA (0, 0, q) = MA (q);

ARIMA (p, 0, q) = ARMA (p,q).

Então, os modelos AR, MA e ARMA usam a série original, supondo estacionariedade.

Alguns operadores úteis no entendimento das notações associadas aos modelos são (Morettin & Toloi, 1987).

(i) operador translação para o passado, denotado por B e definido por:
$BZ(t) = Z(t-1); \quad B^m Z(t) = Z(t-m)$

(ii) operador translação para o futuro, denotado por F e definido por:
$FZ(t) = Z(t+1); \quad F^m Z(t) = Z(t+m)$

(iii) operador diferença: D$Z(t) = Z(t) - Z(t-1) = (1-B)\,Z(t)$, de forma que $\Delta = 1-B$

(iv) operador soma, denotado por S e definido por:

$SZ(t) = \sum_{j=0}^{\infty} Z(t-j) = (1-B)^{-1} Z(t) = \Delta^{-1} Z(t)$, de forma que S = D^{-1}.

Modelos auto-regressivos (AR)

Sendo $Z(t) = Z(t)$ - m, um modelo auto-regressivo de ordem p, denotado por AR(p), é dado por:

$$\tilde{Z}(t) = \phi_1 \tilde{Z}(t-1) + \phi_2 \tilde{Z}(t-2) + \mathrm{K} + \phi_p \tilde{Z}(t-p) + a(t).$$

Definindo-se o operador auto-regressivo estacionário de ordem p:

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \mathrm{K} - \phi_p B^p$, pode-se escrever: $\phi(B)\ \tilde{Z}(t) = a(t).$

O modelo auto-regressivo de ordem p = 1, denominado AR(1) é dado por: $\tilde{Z}(t) = \phi\tilde{Z}(t-1) + a(t)$ e depende apenas de *Z(t-1)* e do ruído branco no instante t.

Alternativamente, tem-se:

$$\tilde{Z}(t) = a(t) + \phi a(t-1) + \phi^2 a(t-2) + \mathrm{K} = \sum_{j=0}^{\infty} \phi^j a(t-j).$$

A função de autocovariância eqüivale a:

$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \mathrm{K} + \phi_p\gamma_{j-p}, \quad j > 0$, a qual, dividida por

$\gamma_0 = Var[Z(t)]$, fornece a função de autocorrelação

$$\rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + \mathrm{K} + \phi_p\rho_{j-p}, \quad j > 0$$

Sendo $Var[\tilde{Z}(t)] = Var[Z(t)] = \gamma_0 = \phi_1\gamma_1 + \mathrm{K} + \phi_p\gamma_{-p} + \sigma_a^2$, obtém-se

$1 = \phi_1\rho_1 + \mathrm{K} + \phi_p\rho_p + \dfrac{\sigma_a^2}{\gamma_0}$, de forma que:

$Var[Z(t)] = Var(Z) = \dfrac{\sigma_a^2}{1 - \phi_1\rho_1 - \mathrm{K} - \phi_p\rho_p}$, em que σ_a^2 é a variância do ruído branco ou variância residual.

Modelos de médias móveis (MA)

Um modelo de médias móveis de ordem q, denotado por MA(q), é dado por: $Z(t) = \mu + a(t) - \theta_1 a(t-1) - \mathrm{K} - \theta_q a(t-q)$.

Sendo $\tilde{Z}(t) = Z(t) - \mu$, tem-se:

$\tilde{Z}(t) = (1 - \theta_1 B - \text{K} - \theta_q B^q) a^t = \theta(B)\ a(t)$, em que

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \text{K} - \theta_q B^q$ é o operador de médias móveis de ordem q.

O modelo de médias móveis de ordem q = 1, denominado MA(1) é dado por:

$\tilde{Z}(t) = a(t) - \theta a(t-1)\ \ ou\ \ \tilde{Z}(t) = (1 - \theta B)\ a(t)$.

As funções de autocovariância e de autocorrelação de um processo (MA(q) são iguais a zero para defasagens ou "lags" maiores do que q, ao contrário do que ocorre com o processo AR (Morettin & Toloi, 1987).

A variância de Z(t) eqüivale a

$Var[Z(t)] = Var(Z) = \gamma_0 = (1 + \theta_1^2 + \text{K} + \theta_q^2)\ \sigma_a^2$ eqüivalendo a

$Var(Z) = (1 + \theta^2)\ \sigma_a^2$ para um processo MA(1).

Modelos auto-regressivos de médias móveis (ARMA)

Os modelos ARMA (p, q) são dados por:

$$\tilde{Z}(t) = \phi_1 \tilde{Z}(t-1) + \text{K} + \phi_p \tilde{Z}(t-p) + a(t) - \theta_1 a(t-1) - \text{K} - \theta_q a(t-q)\ .$$

Sendo $\phi(B)\ e\ \theta(B)$ os operadores auto-regressivos e de médias móveis, respectivamente, pode-se escrever $\phi(B)\ \tilde{Z}(t) = \theta(B)\ a(t)$.

Um modelo ARMA (1,1) é dado por:

$\tilde{Z}(t) = \phi \tilde{Z}(t-1) + a(t) - \theta a(t-1)$ e está associado à variância

$$Var[Z(t)] = Var(Z) = \gamma_0 = \frac{(1 + \theta^2 - 2\phi\theta)}{1 - \phi^2}\ \sigma_a^2\ .$$

Para j > q, as funções de autocovariância e de autocorrelação eqüivalem aquelas dos modelos auto-regressivos.

Modelos auto-regressivos integrados de médias móveis (ARIMA)

Tomando-se um número finito de diferenças d em busca de estacionariedade, tem-se $W(t) = \Delta^d Z(t)$. Sendo $W(t)$ uma diferença de $Z(t)$, então $Z(t)$ é uma integral de $W(t)$, de forma a se dizer que $Z(t)$ segue um modelo auto-regressivo integrado de médias móveis (ARIMA), dado por: $\phi(B)\ \Delta^d Z(t) = \theta(B)\ a(t)$, denominado ARIMA (p, q, d), em que p e q são as ordens de $\phi(B)\ e\ \theta(B)$, respectivamente. Casos particulares deste modelo são:

ARIMA (0, 1, 1): $\Delta Z(t) = (1-\theta B)\ a(t)$, denominado integrado de médias móveis (IMA);

ARIMA (1, 1, 1): $(1-\phi B)\Delta Z(t) = (1-\theta B)\ a(t)$.

Mais detalhes sobre os modelos ARIMA são apresentados no tópico 7.3.

7.3. Modelos ARIMA na análise espacial de experimentos de campo

A modelagem da estrutura de erro como um processo estocástico estacionário foi considerada inicialmente por Martin (1982; 1986). Posteriormente, Gleeson & Cullis (1987) propuseram modelar a tendência nos experimentos de campo como um efeito aleatório representado por um processo ARIMA de baixa ordem, em apenas uma dimensão. Cullis e Gleeson (1991) estenderam esta abordagem para duas dimensões, a qual tem sido usada com grande eficiência (Grondona et al., 1996; Gilmour et al., 1997; Gilmour, 2000). A seguir será apresentada esta abordagem.

Considerando os dados de parcela indexados na ordem de campo, tem-se o seguinte modelo segundo Gleeson & Cullis (1987):

$y = D\tau + \xi + \eta$, em que:

y: vetor de dimensão n contendo as observações;

τ: vetor de dimensão t contendo os efeitos de tratamentos;

D: matriz de incidência para τ, de dimensões (n x t);

ξ: vetor de dimensão n representando os efeitos aleatórios de tendência;

η: vetor de dimensão n representando os efeitos localizados, assumidos como desvios independentes $N(0, \sigma^2)$.

Trocando a notação da série temporal de *Z(t)* para ξ_t e do ruído branco de a(t) para a_t e assumindo que os elementos ξ_t *(t = 1, ..., n)* de ξ do modelo podem ser representados por um processo ARIMA (p, d, q), tem-se na notação de Box & Jenkins (1970):

$\phi(B)\Delta^d\xi_t = \theta(B)\, a_t$, em que:

B: operador de translação para o passado, $B\xi_t = \xi_{t-1}$;

d: nível de diferenciação necessário para fazer $\Delta^d x_t$ estacionário, $\Delta = 1-B$ e

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \mathrm{K} - \phi_p B^p;$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \mathrm{K} - \theta_q B^q;$$

a_t: ruídos brancos assumidos como desvios independentes $N(0,\ \sigma_a^2)$.

Assumindo que os elementos η_t *(t = 1, ..., n)* de η são ruídos brancos adicionais, Gleen & Cullis (1987) mostraram que $\Delta^d(\xi_t + \eta_t)$ é um processo ARMA (p, Q) com θ sendo o maior dentre (p + d) e q. Tais autores relataram também que uma pequena subclasse de modelos ARIMA (p, d, 0) com p = 0 ou 1 e d = 1 ou 2 geralmente propiciam um adequado ajustamento.

Cullis & Gleeson (1991) consideraram um experimento com r linhas e c colunas e denotaram por $W = (w_{ij})$ a matriz dos dados indexados na ordem de campo. Fazendo-se $w = vec(W)$, o modelo para w é dado por:

$w = D\tau + \xi + \eta$, em que:

τ: vetor de dimensão t contendo os efeitos de tratamentos;

ξ: vetor de efeitos aleatórios de tendência, com dimensão $N = rc$;

η: vetor de erros localizados, com dimensão $N = rc$;

D: matriz de incidência para τ com dimensões $(N \times t)$.

O vetor ξ pode ser representado por um processo ARIMA (p_2, d_2, q*) ao longo das linhas. Denotando $\varepsilon_i = \xi_i + \eta_i$ como um erro na parcela (ou planta) i, tem-se na notação de Box e Jenkins para modelos multiplicativos:

$\Phi(B^r)\Delta_r^{d_2}\varepsilon_i = \Theta(B^r)\ \alpha_i$, em que:

$\Delta_r = 1 - B^r$, sendo que B é o operador de translação para o passado;

$\Phi(B^r)\ e\ \Theta(B^r)$: polinômios em B^r de graus p_2 e q_2, respectivamente,

onde $q = máx\ (q^*,\ p_2 + d_2)$.

Este modelo liga os componentes do erro ε_i, r parcelas adiante em w. E se ε_i são correlacionados também nas colunas, os componentes α_i, α_{i-1}, ... serão correlacionados. Para considerar a existência de correlações nas colunas, deve-se introduzir um segundo modelo dado por:

$\phi(B)\Delta^{d_1}\alpha_i = \theta(B)\ a_i$, em que:

a_i : ruído branco

$\phi(B)\ e\ \theta(B)$: polinômios em B de graus p_1 e q_1, respectivamente, em que

$\Delta = 1-B$.

Substituindo o modelo para colunas, no modelo para linhas, tem-se o seguinte modelo geral multiplicativo.

$\phi_{p_1}(B)\ \Phi_{p_2}(B^r)\Delta^{d_1}\Delta_r^{d_2}\varepsilon_i = \theta_{q_1}(B)\ \Theta_{q_2}(B^r)\ a_i$.

Este processo pertence a subclasse de processos látice os quais são denominados **separáveis**. Se $\rho(g,0)$ e $\rho(0, h)$ são as correlações lag-g e lag-h de processos uni-dimensionais ao longo de colunas e linhas, respectivamente, então a **separabilidade** implica

$$\rho(g, h) = \rho(g, 0)\, \rho(0, h)$$

Martin (1979) mostrou que para uma realização de um processo separável em um látice retangular, a matriz de variância eqüivale ao produto de Kronecker das matrizes de variância de processos uni-dimensionais. Assim, para os dados diferenciados, tem-se:

$y = (\Delta_c \otimes \Delta_r)\, w = \Delta w = \Delta D\tau + e = X\tau + e$, em que:

$Var(e) = \sigma^2 H;$

$H = \Sigma_2(\gamma_2) \otimes \Sigma_1(\gamma_1);$

γ_1 e γ_2: vetores $(p_1 + q_1)$ e $(p_2 + q_2)$ de parâmetros dos processos ARMA nas linhas e colunas, ou seja, parâmetros da matriz de covariância;

$X = (\Delta_c \otimes \Delta_r)\, D;$

Δ_r: matriz n_1 x r que promove a diferenciação para a ordem d_1;

Δ_c: matriz n_2 x c que promove a diferenciação para a ordem d_2;

$n_1 = r\text{-}d_1$;

$n_2 = c\text{-}d_2$.

Este modelo geral, desenvolvido por Cullis & Gleeson (1991) é, essencialmente, equivalente ao desenvolvido por Martin (1990).

No contexto dos experimentos de melhoramento, o seguinte modelo linear geral pode ser especificado:

$y = D\tau + Z\beta + S + \eta$, , em que:

y: vetor de dados, ordenados como colunas e linhas dentro de colunas;

τ: vetor representando os efeitos de tratamentos ou de genitores;

β: vetor representando a variação espacial em larga escala ou tendência global (efeitos de blocos, tendência po linomial);

S: representa a variação espacial em pequena escala (dentro de blocos) ou tendência local, modelada como um vetor aleatório com média zero e variância espacialmente dependente;

η: vetor de erros independentes e identicamente distribuídos.

Em termos de processos ARIMA separáveis S pode ser escrito como:

$$\phi_{p_1}(B)\ \Phi_{p_2}(B^r)\Delta^{d_1}\Delta_r^{d_2}S_i = \theta_{q_1}(B)\ \Theta_{q_2}(B^r)\ a_i$$

Usando um modelo deste tipo, Grondona et al. (1996) concluíram que, dentre 19 modelos avaliados em 35 ensaios, aqueles que consideram a correlação espacial em duas dimensões são os mais eficientes, destacando-se o modelo ARIMA (1, 0, 0) x ARIMA (1, 0, 0), ou seja, o modelo AR(1) x AR(1). O modelo AR(1) x AR(1) mostrou-se o mais eficiente em 21 dos 35 ensaios.

No modelo AR(1), em analogia a $Z(t) = \phi Z(t-1) + a(t)$ e sendo $\phi = \rho_1$ a estimativa inicial de ϕ, tem-se que o processo AR(1) padrão para os erros é definido como:

$\xi_i = \rho\xi_{i-1} + a_i$, em que:

ξ_i: erro na posição i;

ξ_{i-1}: erro na posição prévia ou erro correlacionado;

ρ: coeficiente de autocorrelação;

a_i: erro independente na posição i.

A variância do processo AR(1) é dada por $\sigma_\xi^2 = \rho^2\ \sigma_\xi^2 + \sigma_a^2$, a qual rearranjada fornece $\sigma_\xi^2 = \sigma_a^2/(1-\rho^2)$. A covariância de uma série de pontos é dada por $\sigma_\xi^2\left[1\ \rho\ \rho^2\ \rho^3\rho^4 ...\right]$.

Existindo o erro extra independente η, tem-se um novo componente de variância σ_η^2 denominado "nugget variance". Neste caso, a covariância de uma série de pontos passa a ser dada por $\left[\left(\sigma_\xi^2 + \sigma_\eta^2\right)\ \sigma_\xi^2\rho\ \sigma_\xi^2\rho^2\ \sigma_\xi^2\rho^3\ \sigma_\xi^2\rho^4\ \mathrm{K}\right]$.

Sob este novo modelo, a autocorrelação no "lag 1" passa a ser dada $\sigma_\xi^2\rho/(\sigma_\xi^2 + \sigma_\eta^2)$, a autocorrelação no "lag 2" passa a ser dada por $\sigma_\xi^2\rho^2/(\sigma_\xi^2 + \sigma_\eta^2)$, etc. Com a inclusão de η no modelo a ser ajustado, σ_η^2

é estimado à parte e a estimativa da autocorrelação aumenta em magnitude, em relação a autocorrelação estimada com base no modelo sem o ajuste de η. Isto porque, quando σ^2_η é omitida do modelo, a variância correlacionada tem que acomodar toda a variância. E sendo a covariância constante através dos dois modelos, tem-se as seguintes correlações no "lag 1" para os dois modelos:

(i) Modelo sem o ajuste de η

$$\rho_s = \frac{\sigma^2_\xi \ \rho}{\sigma^2_\xi + \sigma^2_\eta}$$

(ii) Modelo com o ajuste de η

$$\rho_c = \frac{\sigma^2_\xi \ \rho}{\sigma^2_\xi} = \rho$$

Verifica-se que ρ_c é maior que ρ_s e que pode ser necessário incluir η no modelo.

8. Modelos Lineares Mistos Espaciais ao Nível de Indivíduos com Estrutura Auto-Regressiva de Erros

8.1. Modelo Geral

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

Um modelo linear misto geral é da forma (Henderson, 1984):

$y = Xb + Za + e$ (**1**), com as seguintes distribuições e estruturas de médias e variâncias:

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

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

em que:

y: vetor de observações;

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

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

e: vetor de erros aleatórios;

G: matriz de variância – covariância dos efeitos aleatórios;

R: matriz de variância – covariância dos erros aleatórios;

0: vetor nulo.

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

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

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

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

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

Quando *G* e *R* não são conhecidas, os componentes de variância a eles

associados podem ser estimados eficientemente empregando-se o procedimento REML (Patterson & Thompson, 1971; Searle et al., 1992). Exceto por uma constante, a função de verossimilhança restrita a ser maximizada, é dada por:

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

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

em que:

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

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

C: matriz dos coeficientes das equações de modelo misto.

A função (*L*) de verossimilhança restrita expressa em termos do logaritmo, pode ser maximizada (visando obter as estimativas REML dos componentes de variância) empregando-se diferentes algoritmos tais quais: (i) "Expectation – Maximization" (EM) de Dempster et al. (1977); (ii) "Derivative Free" (DF) de Graser et al. (1987); (iii) "Average Information" (AI) de Gilmour et al. (1995). Estes algoritmos geraram as denominações EMREML, DFREML e AIREML.

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

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

a: vetor de efeitos genéticos aditivos;

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

σ_a^2: variância genética aditiva;

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

σ_e^2: variância residual.

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

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

$$= Xb + Z_1 a^* + Z_2 p + e, \ em \ que:$$

σ_p^2: variância dos efeitos permanentes.

(c) Modelo multivariado, ajustando os efeitos aditivos

No caso bivariado tem-se:

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

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

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

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

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

(d) Modelo geoestátistico ou de séries temporais para análise espacial

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

Os modelos espaciais (geoestatísticos ou de séries temporais) permitem estudar a variabilidade espacial do solo nas áreas experimentais, através do uso de procedimentos que permitem um melhor critério de estratificação ambiental (para seleção massal ou para melhor definição dos efeitos fixos no procedimento BLUP). A análise espacial pode também ser realizada simultaneamente à predição BLUP dos valores genéticos e estimação REML dos componentes simplesmente incorporando $R = \Sigma$ nos estimadores e preditores.

Assim, as diferenças entre dois modelos, um espacial e outro não, são dadas apenas pela suposição associada ao erro experimental. Neste contexto, é relevante relatar que entender e modelar o erro não é tarefa fácil em Estatística e Genética Quantitativa. Entretanto, é tarefa de suma relevância uma vez que um modelo de erros escolhido aleatoriamente pode não representar a realidade e comprometer todas as inferências.

Dentre as várias formas de modelagem de Σ, o modelo de séries temporais auto-regressivo separável de primeira ordem em duas dimensões (AR1 x AR1) tem-se mostrado superior (Grondona et al., 1996; Gilmour et al., 1997; Cullis et al., 1998; Apiolaza et al., 2000; Qiao et al., 2000; Costa & Silva, 2001). Inclusive é atualmente um dos procedimentos adotados no software ASREML (Gilmour et al., 2000) desenvolvido com o objetivo de permitir análises espaciais de modelos mistos via REML/BLUP e que tem sido usado amplamente (Gilmour et al., 1997; Cullis et al., 1998; Apiolaza et al., 2000; Qiao et al., 2000; Costa & Silva, 2001) para esta finalidade. Assim, tal modelo será detalhado a seguir.

De maneira genérica, o erro é modelado em função de um efeito de tendência (ξ) mais um resíduo aleatório não correlacionado (η). Assim, o vetor de erros é particionado em $e = \xi + \eta$, em que ξ e η referem-se aos vetores de erros espacialmente correlacionados e resíduos aleatórios independentes, respectivamente. Os modelos de análise tradionais não incluem o componente ξ e alguns modelos de análise espacial empregados em agricultura incluem ξ mas ignoram η. O resíduo η contempla efeitos genéticos não aditivos, efeitos de microambiente e erros de medida. A não modelagem de η pode ter como consequência a superestimação da variância genética aditiva. Dessa forma, tal efeito é de capital importância no ajuste de modelos mistos individuais.

Considerando um experimento com forma retangular em uma grade de c colunas e r linhas, os resíduos podem ser arranjados em uma matriz de forma que os mesmos podem ser considerados como correlacionados dentro de linhas e colunas. Escrevendo estes resíduos em um vetor na ordem de campo pela colocação de uma coluna sob outra, a variância dos resíduos é dada por Var(e) = Var (ξ + η) = R = Σ = $\sigma_\xi^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]+I\sigma_\eta^2$, em que σ_ξ^2 é a variância devida a tendência é σ_η^2 a variância dos resíduos não correlacionados. As matrizes $\sum_c(\Phi_c)$ e $\sum_r(\Phi_r)$ referem-se a matrizes de correlação auto-regressivas de primeira ordem com parâmetros de autocorrelação Φ_c e Φ_r e ordem igual ao número de colunas e número de linhas, respectivamente. Assim, ξ é modelado como um processo auto-regressivo separável de primeira ordem (AR1 x AR1) com matriz de covariância $Var(\xi)=\sigma_\xi^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$ (Gilmour et al., 1997; 2000).

Coeficientes de autocorrelação altos (> 0,60) podem indicar padrões de resíduos espaciais ou variabilidade espacial em gradientes ou em manchas. Por outro lado, coeficientes de autocorrelação baixos indicam variabilidade espacial aleatória ou sem padrão definido.

Este modelo preserva a estrutura e informação do delineamento experimental (considera os efeitos de bloco) e pode diferir dos modelos usados em agricultura apenas por considerar σ_η^2 e por ser ao nível de indivíduos e não de parcela. Qiao et al. (2000) relataram a importância de se preservar a informação do delineamento experimental e concluíram que ambos, tanto a informação do delineamento quanto o ajustamento espacial da tendência necessitam ser considerados. Tais autores comentam que bons delineamentos experimentais e apropriados (avançados) procedimentos analíticos são importantes na obtenção de eficiência na seleção.

Tal modelo difere do modelo de análise em blocos casualizados por três parâmetros a mais: σ_ξ^2, Φ_c e Φ_r os quais são estimados por REML. Assim, o

modelo em blocos é hierárquico em relação ao modelo espacial e a significância ou melhoramento do modelo devido a adoção da abordagem espacial pode ser avaliada pelo teste da razão de verossimilhança (LRT) confrontando duas vezes a mudança no log da função de verossimilhança maximizada contra o valor da distribuição qui-quadrado com 3 graus de liberdade, em teste bilateral. As significâncias dos vários efeitos aleatórios ajustados pelo modelo espacial podem também ser avaliadas pelo LRT. Outro critério de seleção do modelo é o "Critério de Informação de Akaike" ou AIC, o qual penaliza a verossimilhança pelo número de parâmetros ajustados independentemente. Segundo este critério, qualquer parâmetro extra (a mais) deve aumentar a verossimilhança em pelo menos uma unidade a fim de que seja incluído no modelo. O AIC é dado por AIC = -2 log L + 2 p, em que L equivale a menos duas vezes o valor do log da função de verossimilhança maximizada e p é o número de parâmetros estimados. Menores valores de AIC refletem um melhor ajustamento global (Akaike, **1974).**

8.2 Software

O software ASREML é o mais indicado para a análise espacial de experimentos via modelos mistos. Para a versão "Windows", o arquivo executável necessário é o ASRWIN.EXE. Para utilização do referido aplicativo na análise de modelos mistos ao nível de plantas individuais, são necessários 3 arquivos: um arquivo de dados, um arquivo de pedigree (com os números de identificação dos genitores precedendo os números de identificação dos descendentes) e um arquivo de comandos (este contendo o modelo de análise, o qual deve ser escrito pelo usuário). Os arquivos devem ser preparados em formato ASCII, sendo que os arquivos de dados e de pedigree podem ser salvos usando o programa Notepad (ou bloco de notas, usando documento texto, ou seja extensão .txt) ou mesmo o EXCEL usando a opção de salvar como "Texto (OS/2 ou MS-DOS), o que produz uma extensão .txt.

O arquivo de comandos deve conter a extensão .as e deve ser escrito com base em 5 seções: (i) linha de título; (ii) definição das colunas do arquivo de dados; (iii) definição dos arquivos de pedigree e de dados; (iv) definição do modelo estatístico; (v) definição do modelo de variância (esta seção não é necessária em alguns casos). Este arquivo com extensão .as pode ser composto e salvo facilmente dentro do próprio ASREML, através da modificação de arquivos pré-existentes. Os arquivos de resultados mais importantes do programa são aqueles

com extensão .sln (o qual apresenta a solução para todos os efeitos do modelo com seus respectivos desvios padrões) e .asr (o qual sumariza os dados e a sequência das iterações, apresenta as estimativas dos componentes de variância, a análise de variância para os efeitos fixos e suas soluções).

Para a análise de variáveis binomiais, considere como exemplo um experimento instalado no delineamento em blocos ao acaso, com 33 progênies de meios irmãos (polinização aberta) e 6 blocos, em que foi avaliada a variável sobrevivência (Sob). Tendo-se salvos os arquivos de dados e de pedigree com os nomes Dados.txt e Pedigree.txt, tem-se o seguinte arquivo com extensão .as:

```
Titulo
 Individuo !P
 Pai
Mae 33
Bloco 6
Sob
Pedigree.txt !SKIP 1 !MAKE
Dados.txt !SKIP 1
Sob !BIN !LOGIT ~ Bloco !r Individuo
```

Este arquivo de comando possui a seguinte interpretação:

(i) Linha 1 : Título qualquer;

(ii) Linhas 2 a 6 : Identificação de colunas no arquivo de dados, contendo os respectivos números de níveis;

(iii) Linhas 7 e 8 : Identificação dos arquivos de pedigree e de dados;

(iv) Linha 9 : Especificação do modelo de análise para a variável Sob;

Ainda no arquivo com extensão .as, tem-se que os comandos advém após o símbolo ! . Assim, tem-se:

P : indica que o número de indivíduos deve ser lido no arquivo de pedigree;

SKIP 1: indica que deve ser ignorada a primeira linha dos arquivos, pois refere-se apenas as identificações;

MAKE : indica que deve ser feita a matriz de parentesco;

BIN : indica que a variável apresenta distribuição binomial;

LOGIT : indica que deverá ser usada a função de ligação logito;

r Individuo: indica que os efeitos de indivíduos são aleatórios;

~ Bloco : indica que os efeitos de blocos são fixos.

No caso, a seção (v) não foi necessária. Outra opção de análise refere-se a adoção de um modelo de genitor, sendo preditos (1/2) dos efeitos genéticos dos genitores e estimado (1/4) da variação genética aditiva. Neste caso, não é necessário o arquivo de pedigree e o arquivo .as apresenta como conteúdo:

```
Titulo
 Individuo 1188
 Mae 33
 Bloco 6
 Sob
Dados.txt !SKIP 1
Sob !BIN !LOGIT ~ Bloco !r Mae
```

O número 1188 refere-se ao número total de indivíduos, considerando que haviam 6 plantas por parcela.

A seguir será ilustrada a aplicação do software para a análise de uma variável contínua, como o peso de frutos em cacau. Considerando a situação de cruzamentos dialélicos com medidas repetidas em cada indivíduo, tem-se o seguinte arquivo .as, considerando cruzamentos envolvendo 5 mães e 5 pais, experimentação em 4 blocos, 4 medições por indivíduo, 10 famílias de irmãos germanos, 80 parcelas e 1200 indivíduos no total:

Titulo

Individuo !P

Pai 5

Mae 5

Bloco 4

Medicao 4

Familia 10

Permanente 1200

Parcela 80

Peso

Pedigree.txt !SKIP 1 !MAKE !REPEAT

Dados.txt !SKIP 1

Peso ~ Bloco Medicao !r Individuo Familia Permanente Parcela

No caso, os efeitos de bloco e medição foram ajustados como fixos. O comando REPEAT indica que trata-se de um caso de medidas repetidas.

Após compostos os arquivos de dados, de pedigree e de comandos, basta executar este último e, então, abrir os arquivos de resultados.

Considere a avaliação da variável número de frutos (denominada Frutos) de cacaueiros em dois locais, no delineamento em blocos ao acaso com 18 progênies de meios irmãos e cinco blocos em cada local. Assumindo normalidade, tem-se que a composição do arquivo de comandos .as para a análise do modelo bivariado, eqüivale a:

Titulo

Individuo !P

Pai

Mae 18

Bloco 10

Parcela 180

Frutos1 !M 0

Frutos2 !M 0

Pedigree.txt !SKIP 1 !MAKE

Dados.txt !SKIP 1

!ASUV

Frutos1 Frutos2 ~ Trait Tr.Bloco !r Tr.Individuo Tr.Parcela !f mv 2 1 2

0 !$s_2 = ve_1$

0 !$s_2 = ve_2$

Tr.Individuo 2

Tr 0 US va_1 va_{12} va_2

Individuo

Tr.Parcela 2

Tr 0 US vc_1 vc_2

Parcela

!end

O comando !M0 deve ser incluído visando converter os valores zero do arquivo em valores inexistentes ou perdidos. No arquivo, os dados referentes a cada local devem ser colocados em duas colunas distintas, prenchendo-se os dados inexistentes com zero (neste caso).

Por outro lado, o comando !ASUV é usado quando os dados são apresentados em uma forma multivariada mas a análise requerida refere-se a um único caráter. Com esta opção, se existe valores perdidos no arquivo de dados, deve-se incluir o comando !f mv no final da linha do modelo linear. O comando !ASUV deve ser colocado em uma linha logo após a linha de denominação do arquivo de dados e antes da linha referente ao modelo linear.

As linhas após o modelo linear referem-se a definição do modelo (estrutura) de variância. No caso, ve_1 e ve_2 referem-se aos valores iniciais para as variâncias residuais nos locais 1 e 2, respectivamente, e vc_1 e vc_2 referem-se aos valores iniciais para as variâncias entre parcelas nos locais 1 e 2, respectivamente. Por outro lado, va_1, va_2 e va_{12}, referem-se aos valores iniciais para as variâncias genéticas aditivas nos locais 1 e 2 e covariância genética aditiva entre os locais 1 e 2, respectivamente. Assim, tais valores numéricos devem ser fornecidos na estruturação do arquivo de comandos.

Para ajuste de modelos espaciais, deve-se declarar a variância associada a estrutura de correlação através dos comandos AR, MA ou ARMA, referindo-se aos modelos auto-regressivos, médias móveis ou auto-regressivo de médias móveis, respectivamente. O comando ! units declara a inclusão do termo de erro não correlacionado.

Considerando o exemplo com 33 progênies de meios irmãos, 6 blocos e 6 plantas por parcela, tem-se o seguinte arquivo de comandos (ignorando o efeito de parcela) para o ajuste de um modelo espacial auto-regressivo de primeira ordem em duas dimensões com um termo de erro não correlacionado:

```
Título
Individuo !P
Pai
Mae 33
Bloco 6
Coluna 33
 Linha 36
 Sob
Pedigree.txt !SKIP 1 !MAKE
```

Dados.txt !SKIP 1

Sob !BIN !LOGIT ~ Bloco !r Individuo ! units

1 2

Coluna Coluna AR .3

Linha Linha AR .9

O comando 1 2 declara que o erro correlacionado deve ser modelado como um produto direto separável de 2 termos, ou seja, como

$$Var(\xi) = \sigma^2_{\xi}[\sum_{c}(\Phi_c) \otimes \sum_{r}(\Phi_r)] .$$

O comando Coluna Coluna AR .3 indica que o primeiro componente da estrutura de correlação é definido pelo fator coluna com uma matriz de correlação auto-regressiva com um valor inicial do parâmetro de correlação igual a 0,3. O segundo termo Coluna declara para o ASREML ordenar os dados, uma vez que tais podem não estar na ordem de campo.

O comando Linha Linha AR .9 indica que o segundo componente da estrutura de correlação é definido pelo fator linha com uma matriz de correlação auto-regressiva com um valor inicial do parâmetro de correlação igual a 0,9.

Em geral, a seguinte sequência de modelos pode ser ajustada:

(a) Modelo de blocos ao acaso (Modelo1)

Titulo

Individuo 84

Clone 14

Bloco 2

Parcela 28

Peso

Dados.txt !SKIP 1

Peso ~ Bloco !r Clone Parcela

(b) Modelo (AR1 xAR1) sem a inclusão do termo de erro independente (Modelo 2)

```
Titulo
Individuo 84
Clone 14
Bloco 2
Parcela 28
Coluna 14
Linha 6
Peso
Dados.txt !SKIP 1
Peso ~ Bloco !r Clone Parcela
1 2
Coluna Coluna AR .3
Linha Linha AR .9
```

(c) Modelo (AR1 xAR1) com a inclusão do termo de erro independente (Modelo 3)

Quando ambas as correlações auto-regressivas são significativas, o termo de erro independente pode ser acrescentado visando verificar se ocorrem melhorias no ajuste do modelo. O programa é então, dado por:

```
Titulo
Individuo 84
Clone 14
Bloco 2
Parcela 28
Coluna 14
```

Linha 6

Peso

Dados.txt !SKIP 1

Peso ~ Bloco !r Clone Parcela units

1 2

Coluna Coluna AR .3

Linha Linha AR .9

(d) Modelo (AR1 xAR1) com a inclusão de linhas e colunas como efeitos aleatórios (Modelo 4)

Titulo

Individuo 84

Clone 14

Bloco 2

Parcela 28

Coluna 14

Linha 6

Peso

Dados.txt !SKIP 1

Peso ~ Bloco !r Clone Parcela Coluna Linha

1 2

Coluna Coluna AR .3

Linha Linha AR .9

Algumas vezes os efeitos de colunas e linhas mostram-se significativos devido a práticas diferenciadas de cultura e de colheita.

8.3. Aplicação a dados experimentais

Foram considerados dados referentes a um experimento de avaliação de populações, progênies e indivíduos de erva-mate para dois caracteres: massa

foliar em poda de formação (peso1) e massa foliar em poda de produção (peso2). O experimento contemplou 8 populações e 141 progênies avaliadas em 10 blocos casualizados com 6 plantas por parcela, perfazendo um total de 8460 indivíduos. Para cada caráter foram considerados dois conjuntos de dados, um contemplando os 10 blocos e outro usando apenas os 3 primeiros blocos.

Os resultados referentes ao conjunto completo de dados são apresentados nas Tabelas 1 (considerando o efeito de blocos como fixo) e Tabela 2 (considerando o efeito de blocos como aleatório). Os resultados referentes ao conjunto de 3 blocos são apresentados na Tabelas 3 e 4.

Tabela 1. Estimativas dos parâmetros variância genética entre progênies ($\hat{\sigma}^2_g$), entre procedências ($\hat{\sigma}^2_{proc.}$), entre parcelas ($\hat{\sigma}^2_c$), residual ($\hat{\sigma}^2_e$), variância fenotípica ($\hat{\sigma}^2_{yi}$), herdabilidade no sentido restrito ($\hat{h}^2_{ai}$), herdabilidade individual entre procedências ($\hat{h}^2_{proc.}$), coeficiente de determinação dos efeitos de parcela ($\hat{c}^2$), coefientes de autocorrelação residual nas colunas (AR Coluna) e linhas (AR Linha). Dados do experimento completo.

Parâmetros	Peso 2		Peso 1	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
$\hat{\sigma}^2_g$	0,08344	0,09396	0,01107	0,01261
$\hat{\sigma}^2_{proc.}$	0,4167	0,4276	0,04428	0,04164
$\hat{\sigma}^2_c$	0,1385	0,0178	0,03654	0,003211
$\hat{\sigma}^2_e$	1,3390	1,4416	0,2214	0,2479
$\hat{h}^2_{ai}$	0,1688	0,1900	0,1413	0,1652
$\hat{h}^2_{proc.}$	0,2107	0,2160	0,1413	0,1364
$\hat{c}^2$	0,07	0,00	0,12	0,01
Log L	-5926,86	-5712,26	1552,89	1931,14
$\hat{\sigma}^2_{yi}$	1,9776	1,9794	0,3133	0,3054
AR Coluna	---	0,2057	----	0,2877
AR Linha	---	0,1259	----	0,1399
Mudança em Log L	---	214,6**	----	378,25**
Eficiência da análise espacial ($\hat{h}^2$)		1,13	-----	1,17
$\hat{\rho}_{gf} = (1/4)\hat{h}^2_{ai} / [(1/4)\hat{h}^2_{ai} + \hat{c}^2]$	0,376	------	0,2325	-----
$\hat{c}^2 / \hat{h}^2_{ai}$	0,41	------	0,85	-----
$\hat{\rho}_{gi} = \hat{h}^2_{ai} / (\hat{h}^2_{ai} + \hat{c}^2)$	0,71	------	0,54	-----

Tabela 2. Estimativas dos parâmetros variância genética entre progênies ($\hat{\sigma}_g^2$), entre procedências ($\hat{\sigma}_{proc.}^2$), entre parcelas ($\hat{\sigma}_c^2$), entre blocos ($\hat{\sigma}_b^2$), residual ($\hat{\sigma}_e^2$), variância fenotípica ($\hat{\sigma}_{yi}^2$), herdabilidade no sentido restrito ($\hat{h}_{ai}^2$), coeficiente de determinação dos efeitos de parcela ($\hat{c}^2$), e de blocos ($\hat{b}^2$), coefientes de autocorrelação residual nas colunas (AR Coluna) e linhas (AR Linha). Dados do experimento completo.

Parâmetros	Peso 2		Peso 1	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
$\hat{\sigma}_g^2$	0,08344	0,09397	0,01107	0,01261
$\hat{\sigma}_{proc.}^2$	0,4167	0,4277	0,04428	0,04164
$\hat{\sigma}_c^2$	0,1385	0,0178	0,03654	0,003211
$\hat{\sigma}_b^2$	0,03957	0,03804	0,007964	0,003183
$\hat{\sigma}_e^2$	1,3389	1,4417	0,2214	0,2479
$\hat{h}_{ai}^2$	0,1654	0,1862	0,1378	0,1613
$\hat{c}^2$	0,07	0,00	0,115	0,01
$\hat{b}^2$	0,02	0,02	0,025	0,025
Log L	-5918,26	-5703,58	1568,70	1947,17
$\hat{\sigma}_{yi}^2$	2,0171	2,0192	0,3213	0,3127
AR Coluna	---	0,2057	----	0,2877
AR Linha	---	0,1257	----	0,1401
Mudança em Log L	---	214,68*	----	378,47**
Eficiência da análise espacial ($\hat{h}^2$)	---	1,13	----	1,17
$\hat{\rho}_{gf} = (1/4)\hat{h}_{ai}^2 / [(1/4)\hat{h}_{ai}^2 + \hat{c}^2]$	0,37	------	0,23	-----
$\hat{c}^2 / \hat{h}_{ai}^2$	0,42	------	0,83	-----
$\hat{\rho}_{gi} = \hat{h}_{ai}^2 / (\hat{h}_{ai}^2 + \hat{c}^2)$	0,70	------	0,545	-----

Verificou-se que o modelo espacial (modelo 2) apresentou melhor ajuste que o modelo tradicional em blocos casualizados (modelo 1) para ambos os caracteres, conforme pode ser visto pelas mudanças em Log L, as quais foram altamente significativas pelo teste qui-quadrado. Constataram-se, também, aumentos das estimativas da variância genética entre progênies e do coeficiente de herdabilidade e consequentemente das eficiências seletivas que equivaleram a 1,13 e 1,17 para os caracteres peso 2 e peso 1, respectivamente (Tabelas 1 e 2).

A principal diferença entre as estimativas propiciadas pelos 2 modelos referem-se à variância ambiental entre parcelas e consequentemente ao componente c^2. No modelo 1, as estimativas de c^2 foram da ordem de 0,07 e 0,12 ao passo que sob o modelo 2 estas estimativas caíram a praticamente 0% (Tabelas 1 e 2). Em outras palavras, a vantagem do modelo espacial advém da redistribuição da variação ambiental entre parcelas, na variância genética e na variância residual. Sob o modelo de blocos aleatórios (Tabela 2), verificou-se que a análise espacial não contribuiu para a redução da variação entre blocos e do coeficiente de determinação dos efeitos de bloco (b^2), mostrando que o modelo auto-regressivo não se mostrou capaz de remover a variação entre blocos. Este resultado, revela a importância do delineamento experimental e da consideração da informação deste delineamento por ocasião da análise estatística.

Pode-se inferir que a análise espacial, via modelos auto-regressivos propicia, por análise, o que o experimentador deseja, ou seja, blocos homogêneos sem variação ambiental entre parcelas dentro deles. A vantagem do uso deste procedimento analítico está diretamente relacionado à magnitude do coeficiente c^2 e, sobretudo, da magnitude deste em relação a herdabilidade individual, a qual pode ser avaliada por c^2/h^2 ou pelo próprio ρ_g que, no caso de testes de progênies de polinização aberta, pode ser dado por $\rho_{gi} = h_{ai}^2/(h_{ai}^2 + c^2)$ ou $\rho_{gf} = (1/4)h_{ai}^2/[(1/4)h_{ai}^2 + c^2]$ aos níveis de indivíduo e de famílias, respectivamente. Com base nos resultados das Tabela 1; 2; 3 e 4 pode-se inferir que a análise espacial trará vantagens principalmente quando $c^2/h^2 \geq (1/3)$, ou equivalentemente, $\rho_{gi} \leq 0{,}75$ ou $\rho_{gf} \leq 0{,}43$. Isto equivale a c^2 de magnitudes iguais ou superiores a 0,05; 0,07 e 0,10 para herdabilidades individuais da ordem de 0,15; 0,20 e 0,30, respectivamente.

Tabela 3. Estimativas dos parâmetros variância genética entre progênies ($\hat{\sigma}^2_g$), entre procedências ($\hat{\sigma}^2_{proc.}$), entre parcelas ($\hat{\sigma}^2_c$), residual ($\hat{\sigma}^2_e$), do erro independente ($\hat{\sigma}^2_\eta$), variância fenotípica ($\hat{\sigma}^2_{yi}$), herdabilidade no sentido restrito ($\hat{h}^2_{ai}$), herdabilidade individual entre procedências ($\hat{h}^2_{proc.}$), coeficiente de determinação dos efeitos de parcela ($\hat{c}^2$), variância entre colunas ($\hat{\sigma}^2_{col.}$), variância entre linhas ($\hat{\sigma}^2_{linha}$), coefientes de autocorrelação residual nas colunas (AR Coluna) e linhas (AR Linha). Dados de apenas três repetições.

Parâmetros	Peso2			
	Modelo 1	Modelo 2	Modelo 3	Modelo 4
$\hat{\sigma}^2_g$	0,022107	0,054938	0,066147	0,06022
$\hat{\sigma}^2_{proc.}$	0,4652	0,4632	0,4396	0,4625
$\hat{\sigma}^2_c$	0,2079	0,0000	0,0000	0,000
$\hat{\sigma}^2_e$	1,42601	1,5819	0,5768	1,5696
$\hat{\sigma}^2_\eta$	-----	-----	1,0485	-----
$\hat{h}^2_{ai}$	0,0417	0,1046	0,1242	0,1151
$\hat{h}^2_{proc.}$	0,2193	0,2206	0,2063	0,2210
$\hat{c}^2$	0,098	0,00	0,00	0,00
Log L	-1881,54	-1803,23	-1750,05	-1801,03
$\hat{\sigma}^2_{yi}$	2,1212	2,1000	2,1310	2,0923
AR Coluna	-----	0,2186	0,7883	0,2120
AR Linha	-----	0,1698	0,5657	0,1701
Mudança em Log L	-----	78,31**	131,49**	80,51**
Eficiência da análise espacial ($\hat{h}^2$)	-----	2,51	2,98	2,76
$\hat{\sigma}^2_{col.}$	-----	-----	-----	0,000
$\hat{\sigma}^2_{linha}$	------	-----	-----	0,01641
$s = \hat{\sigma}^2_e + \hat{\sigma}^2_\eta$	------	-----	1,6253	-----
$s = \hat{\sigma}^2_e + \hat{\sigma}^2_c$	1,6339	1,5819	-----	1,5696
$s/(\sigma^2_e + \sigma^2_c)$ *Modelo* 1	------	0,9682	0,9947	0,9606
$\hat{\rho}_{gf} = (1/4)\hat{h}^2_{ai}/[(1/4)\hat{h}^2_{ai} + \hat{c}^2]$	0,096	------	-----	-----
$\hat{c}^2 / \hat{h}^2_{ai}$	2,35	------	-----	-----
$\hat{\rho}_{gi} = \hat{h}^2_{ai}/(\hat{h}^2_{ai} + \hat{c}^2)$	0,30	------	-----	-----

Tabela 4. Estimativas dos parâmetros variância genética entre progênies ($\hat{\sigma}_g^2$), entre procedências ($\hat{\sigma}_{proc.}^2$), entre parcelas ($\hat{\sigma}_c^2$), residual ($\hat{\sigma}_e^2$), do erro independente ($\hat{\sigma}_\eta^2$), variância fenotípica ($\hat{\sigma}_{yi}^2$), herdabilidade no sentido restrito ($\hat{h}_{ai}^2$), herdabilidade individual entre procedências ($\hat{h}_{proc.}^2$), coeficiente de determinação dos efeitos de parcela ($\hat{c}^2$), variância entre colunas ($\hat{\sigma}_{col.}^2$), variância entre linhas ($\hat{\sigma}_{linha}^2$), coefientes de autocorrelação residual nas colunas (AR Coluna) e linhas (AR Linha). Dados de apenas três repetições.

Parâmetros	Peso1			
	Modelo 1	Modelo 2	Modelo 3	Modelo 4
$\hat{\sigma}_g^2$	0,008516	0,01147	0,01002	0,01139
$\hat{\sigma}_{proc.}^2$	0,04439	0,04062	0,03950	0,04089
$\hat{\sigma}_c^2$	0,0326	0,0000	0,0000	0,0000
$\hat{\sigma}_e^2$	0,23865	0,2623	0,1181	0,2581
$\hat{\sigma}_\eta^2$	------	-------	0,1507	-------
$\hat{h}_{ai}^2$	0,1051	0,1459	0,1259	0,1468
$\hat{h}_{proc.}^2$	0,1369	0,1292	0,1241	0,1317
$\hat{c}^2$	0,1006	0,00	0,00	0,00
Log L	381,369	512,533	564,110	516,421
$\hat{\sigma}_{yi}^2$	0,3242	0,3144	0,3183	0,3104
AR Coluna	------	0,2874	0,70303	0,2768
AR Linha	------	0,1632	0,5059	0,1594
Mudança em Log L	------	131,16**	182,74**	135,05**
Eficiência da análise espacial ($\hat{h}^2$)	------	1,39	1,20	1,40
$\hat{\sigma}_{col.}^2$	------	-------	------	0,0000
$\hat{\sigma}_{linha}^2$	-----	-------	------	0,0048
$s = \hat{\sigma}_e^2 + \hat{\sigma}_\eta^2$	------	.....	0,2688	-------
$s = \hat{\sigma}_e^2 + \hat{\sigma}_c^2$	0,2713	0,2623	-------	0,2581
$s/(\sigma_e^2 + \sigma_c^2)\,Modelo_1$	1,00	0,9668	0,9907	0,9513
$\hat{\rho}_{gf} = (1/4)\hat{h}_{ai}^2/[(1/4)\hat{h}_{ai}^2 + \hat{c}^2]$	0,207	------	-------	-----
$\hat{c}^2 / \hat{h}_{ai}^2$	0,957	------	------	-----
$\hat{\rho}_{gi} = \hat{h}_{ai}^2/(\hat{h}_{ai}^2 + \hat{c}^2)$	0,51	------	-------	-----

Os resultados das Tabelas 3 e 4 revelam que os modelos 2, 3 e 4 apresentaram ajuste significativamente melhor que o modelo 1, conforme pode ser visto pela mudança em Log L. Por outro lado, as diferenças entre os ajustes dos modelos 2 e 4 tenderam a ser não significativas, mostrando que os efeitos de linhas e colunas apresentaram pequena magnitude conforme pode ser observado por $\hat{\sigma}^2_{col.}$ e $\hat{\sigma}^2_{linha}$ nas Tabelas 3 e 4. As diferenças de ajuste entre os modelos 2 e 3 foram altamente significativas revelando a necessidade de ajuste do termo de erro independente η.

Todos os três modelos de análise espacial absorveram a variação ambiental entre parcelas, tornando os valores de $\hat{C}^2$ iguais a zero. Verifica-se, também, que a soma das variâncias residuais $(\hat{\sigma}^2_e + \hat{\sigma}^2_\eta)$ *ou* $(\hat{\sigma}^2_e + \hat{\sigma}^2_c)$ foram sempre menores pelos modelos 2, 3 e 4 em relação ao modelo 1, mostrando que a abordagem espacial transformou parte da variação residual pelo modelo 1 em variância genética pelos modelos 2, 3 e 4 (Tabelas 3 e 4).

Em geral, o modelo de melhor ajuste foi o 3, o qual elevou a herdabilidade estimada de 0,04 para 0,12 e de 0,105 para 0,126 para os caracteres peso 2 e peso 1, respectivamente, com eficiências seletivas de 2,98 e 1,20 para peso 2 e peso 1, respectivamente (Tabelas 3 e 4). É importante relatar que o modelo de melhor ajuste não é, necessariamente, aquele que fornece maior estimativa de herdabilidade, mas aquele que fornece, significativamente, maior Log L.

As estimativas dos coeficientes de autocorrelação serial foram todas positivas (Tabelas 1; 2, 3 e 4) indicando uma correlação positiva entre o crescimento de plantas vizinhas devido à similaridade ambiental. Estimativas positivas eram mesmo esperadas em erva-mate, tendo em vista que o sistema de utilização da cultura (podas bianuais) minimiza os efeitos de competição, os quais tendem a gerar correlações negativas entre o crescimento de plantas vizinhas. Em espécies florestais, a competição tende a ocorrer em idades avançadas, fato que pode impactar (diminuir a magnitude) as estimativas de autocorrelação e reduzir a eficiência da análise espacial quando não se contemplar a competição nos modelos. As estimativas de autocorrelação aumentaram do modelo 2 para o modelo 3, uma vez que este último modelo ajusta o termo de erro independente e, consequentemente, diminui a variância correlacionada.

Outro fator interessante observado refere-se à obtenção das estimativas de herdabilidade quando foram usados os dados completos e apenas 3 repetições. Pelo modelo em blocos casualizados as estimativas de herdabilidade com o dado completo equivaleram a 0,17 e 0,14 para peso 2 e peso 1, respectivamente, e com apenas 3 repetições equivaleram a 0,04 e 0,105 para peso 2 e peso 1, respectivamente. A análise espacial dos dados de 3 repetições pelo modelo 3 conduziram a estimativas de herdabilidade equivalentes a 0,124 e 0,126, portanto, bem mais próximas às estimadas com os dados completos. Isto mostra que a análise espacial pode melhorar as estimativas e minimizar um pouco os efeitos adversos da amostragem restrita, em algumas situações.

Costa e Silva et al.(2001) realizaram análises espaciais de 12 experimentos com espécies florestais (*Pinus* e *Picea*) empregando o software ASREML. Concluíram que houveram melhoras no modelo com a inclusão dos parâmetros espaciais e que as estimativas dos parâmetros de autocorrelação apresentaram alta magnitude (as estimativas de Φ_c e Φ_r foram superiores a 0,60). Também, Kusnadar & Galwey (2000) constataram que um modelo espacial propiciou melhor ajuste que o delineamento de blocos incompletos para dados de crescimento individual em *Pinus pinaster.*

A abordagem espacial tende a ser vantajosa mas a sua eficiência não pode ser generalizada. Esta eficiência depende de cada situação experimental e do caráter avaliado. Recomenda-se realizar, previamente, uma análise ambiental especulativa (tópico 3) antes de se enveredar pelas aplicações de modelos complexos. Mas tal abordagem é relevante e deve ser difundida, pois, na pior das hipóteses produz resultados idênticos aos da análise tradicional.

9. Análise de dados longitudinais no melhoramento

Os caracteres de interesse no melhoramento de plantas se expressam mais de uma vez em cada indivíduo, gerando dados longitunais ou medidas repetidas. Tais caracteres são denominados infinitamente dimensionais. Dados destes caracteres apresentam estrutura correlacionada através do tempo, safras ou medições. Algumas alternativas de análise destes caracteres são:

(a) Análise univariada considerando cada idade ou safra em separado (Estrutura de covariância não correlacionada).

Neste caso, as avaliações em diferentes estágios são consideradas como sendo caracteres diferentes e não correlacionados. Assim, os dados de cada idade são analisados separadamente. Nesta situação, a estrutura geral (válida para efeitos genéticos e ambientais) de covariância entre quatro diferentes idades ou safras é dada por:

$$\sum = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

A estrutura de covariância genética é dada por:

$$\sum_g = \begin{bmatrix} \sigma^2_{g_1} & 0 & 0 & 0 \\ 0 & \sigma^2_{g_2} & 0 & 0 \\ 0 & 0 & \sigma^2_{g_3} & 0 \\ 0 & 0 & 0 & \sigma^2_{g_4} \end{bmatrix}$$, em que a $\sigma^2_{g_i}$ é a variância genética na idade i.

Este modelo não é realista e deve ser evitado. Em algumas plantas perenes, uma só avaliação as vezes é considerada. Por exemplo, em espécies florestais como o eucalipto, a seleção precoce é enfatizada procurando-se uma idade mínima que torne a seleção eficiente. Em geral, a maioria dos programas conduzidos com esta espécie no Brasil utilizam eficientemente a seleção baseada em uma só avaliação aos três anos.

Em outras plantas perenes coma as fruteiras, muitas vezes são gerados apenas um dado por indivíduo, dado este referente a soma ou média de várias safras. A seleção baseada apenas neste dado totalizado ou médio raramente será completamente eficiente, a menos que todos os indivíduos tenham apresentado

produção em todas as safras. Por outro lado, a seleção baseada em médias por grupo de plantas e em várias safras, como a seleção de genitores baseada na progênie, a seleção de clones em testes clonais e a seleção de híbridos ou famílias híbridas, tende a ser menos prejudicada pelo uso de um único dado médio.

(b) Análise pelo modelo de repetibilidade (estrutura de covariância completamente correlacionada)

Nesta situação, considera-se que o caráter é o mesmo através das várias safras. Isto implica assumir que a correlação genética entre qualquer par de safras equivale a 1, que a variância fenotípica é a mesma para todas as safras e que a correlação ambiental permanente é a mesma para todos os pares de safras. Estas premissas são atendidas aproximadamente para algumas fruteiras mas não para todas. Todo o conjunto de dados é analisado simultaneamente e a estrutura de covariância genética entre quatro diferentes safras é dada por:

$$\sum_g = \sigma_g^2 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$, em que σ_g^2 é a variância genética.

A estrutura geral de covariância entre quatro diferentes safras é dada por:

$$\sum = \begin{bmatrix} k & a & a & a \\ a & k & a & a \\ a & a & k & a \\ a & a & a & k \end{bmatrix},$$

em que K representa os elementos diagonais (variâncias) constantes e a representa os valores de correlação entre pares de safras ou idade.

Em espécies florestais, este modelo de análise é raramente aplicável. Isto porque, com o aumento dos caracteres de crescimento com a idade, as suposições de igual variância e correlação não são realistas. Neste caso, a padronização prévia dos dados remove apenas parcialmente a heterogeneidade de variâncias.

(c) **Modelo multivariado** (Matriz de Covariância não Estruturada)

Este é o modelo mais completo, o qual utiliza toda a informação simultaneamente e trata idades ou safras como sendo caracteres diferentes e correlacionados, considerando suas diferentes herdabilidades e correlações genéticas. A estrutura geral de covariância entre quatro diferentes idades é dada por:

$$\sum = \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix}$$, em que *a, b, c, d, e* e *f,* representam diferentes valores de correlação.

Neste caso, a matriz Σ é denomida não estruturada e o modelo é superparametrizado e proibitivo na prática quando muitas idades ou safras são consideradas. Sob modelo multivariado surgem problemas de estimação (matrizes não positivas definidas) principalmente quando o valor paramétrico encontra-se próximo ao limite do espaço do parâmetro.

(d) **Modelo de Regressão Aleatória** (Estrutura de covariância definida por funções de covariância)

O modelo regressão aleatória pode ser considerado como um modelo multivariado reduzido e simplificado, o qual permite a obtenção dos mesmos parâmetros de interesse (herdabilidade em cada idade e correlação genética entre todos os pares de idade), que podem ser obtidos pelo modelo multivariado, porém com uma menor parametrização e com menor esforço computacional. Tal abordagem define diretamente funções de covariância contínuas e permite incluir na análise indivíduos com idades heterogêneas. A estrutura geral de covarância entre quatro diferentes idades é dada por:

$$\sum = \begin{bmatrix} z_{01} & z_{11} & z_{21} & z_{31} \\ z_{02} & z_{12} & z_{22} & z_{32} \\ z_{03} & z_{13} & z_{23} & z_{33} \\ z_{04} & z_{14} & z_{24} & z_{34} \end{bmatrix} \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix} \begin{bmatrix} z_{01} & z_{02} & z_{03} & z_{04} \\ z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ z_{31} & z_{32} & z_{33} & z_{34} \end{bmatrix}$$

$= Q_i \Lambda_0 Q_i$, em que:

z_{ij} = i-ésimo vetor polinomial ortogonal analisado na idade variável j ;

Λ_0 **=** matriz de covariância dos regressores aleatórios;

Qi = matriz com q + 1 colunas contendo z_0, z_1, z_2, ..., z_q, respectivamente, em que *q* é a ordem do polinômio e z_1 é o i-ésimo vetor polinomial ortogonal.

Um modelo de regressão aleatória pode ser ajustado somente com um intercepto e inclinação, com um termo quadrático adicional, com um termo cúbico adicional e assim sucessivamente de acordo com a ordem do polinômio de Legendre ajustado.

(e) Modelo Auto–Regressivo com Variâncias Heterogêneas (Estrutura Auto – Regressiva)

Este modelo assume correlações diferentes entre idades e reconhece a existência de correlação serial entre as medidas repetidas. A estrutura geral de covarância envolvendo quatro diferentes safras é definida como:

$$\sum = \begin{bmatrix} 1 & a^{|t_2 - t_1|} & a^{|t_3 - t1|} & a^{|t_4 - t_1|} \\ a^{|t_2 - t_1|} & 1 & a^{|t_3 - t_2|} & a^{|t_4 - t_2|} \\ a^{|t_3 - t_1|} & a^{|t_3 - t_2|} & 1 & a^{|t_4 - t_3|} \\ a^{|t_4 - t_1|} & a^{|t_4 - t_2|} & a^{|t_4 - t_3|} & 1 \end{bmatrix}$$

Verifica-se que o modelo *AR* estima uma só correlação e projeta-a para os demais *lags*. Neste caso, a matriz de covariância genética é dada por G = S Σ S em que S é uma matriz diagonal com elementos equivalentes à raiz quadrada da variância genética em cada idade.

(f) Modelo com correlações específicas para cada intervalo de idade (Estrutura de Covariância Toeplitz)

Este modelo considera correlações iguais para iguais intervalos entre idades e variâncias heterogêneas entre idades. É denominado modelo de correlação bandada e a matriz de covariância entre idades tem estrutura Toeplitz, dada por:

$$\sum = \begin{bmatrix} 1 & a & b & c \\ a & 1 & a & b \\ b & a & 1 & a \\ c & b & a & 1 \end{bmatrix}$$

Uma matriz Toeplitz apresenta várias diagonais (primárias, secundárias, terciárias, etc). A matriz de covariância genética é dada por G = S Σ S, em que Σ no caso, é uma matriz Toeplitz.

(g) Ajuste de uma curva spline cúbica no intervalo de idades considerado

Uma spline cúbica é uma curva alisada ou suavizada sobre um intervalo [a,b] formado pela ligação de segmentos polinomiais cúbicos a "pontos nó" tais que a curva inteira e suas primeiras e segundas diferenças são contínuas sobre o intervalo. Refere-se a um processo de interpolação que propicia uma curva contínua e suave a partir dos pontos observados, sendo diferenciável e integrável para um domínio de interesse.

A função spline cúbica usa um processo polinomial de terceiro grau para a interpolação de valores entre cada par de pontos observados. Neste caso, uma polinomial diferente é usada para cada intervalo e cada uma é construida de forma a sempre passar pelos dados originais e apresentar uma derivada contínua nas junções entre cada intervalo. Uma curva definida por uma polinomial cúbica pode passar exatamente por quatro pontos, sendo que, no caso de uma longa sequência de pontos, torna-se necessário usar uma sucessão de intervalos polinomiais. Para garantir que não hajam mudanças súbitas nas inclinações ou curvaturas entre os intervalos sucessivos, tal função polinomial não é ajustada para quatro pontos e sim para dois.

A função spline cúbica foi desenvolvida inicialmente por Shoenberg (1946) e aperfiçoada por Walsh et al. (1962). Verbyla et al. (1999) e White et al. (1999)

sugeriram o uso de funções mais flexíveis tais quais polinomiais de elevada ordem ou splines cúbicas para modelar dados longitunais quando não existe conhecimento prévio sobre o modelo biológico adjacente ao caráter.

Exemplos e detalhes da obtenção de splines cúbicas são apresentados por Davis (1986) e Green & Silverman (1994).

Uma comparação entre as quatro primeiras abordagens (estruturas (a) até (d)) na estimação de parâmetros genéticos em *E. urophylla* avaliado em 7 idades (1 a 7 anos), revelaram a adequabilidade da técnica de regressão aleatória para modelagem do crescimento (Resende, 1999; Resende et al. 2001). Comparação similar foi realizada por Apiolaza et al. (2000) em *Pinus radiata*, incluindo também as estruturas auto-regressivas (AR) e Toeplitz, as quais impõem fortes restrições às covariâncias gerando também modelos mais parcimoniosos, assim como a regressão aleatória. Tais autores relataram que os modelos de regressão aleatória não são necessariamente adequados para qualquer conjunto de dados, tendo verificado que os modelos simples, como o AR, com variâncias específicas para cada idade, parecem ser flexíveis o suficiente para serem usados com eficiência em muitas situações no melhoramento florestal. Pelo modelo *AR*, o valor genético de um indivíduo i na época j, é dado por $a_{ij} = \rho^{[j-(j-1)]} a_{ij-1} + \alpha_j$ em que $\boldsymbol{a_{ij-1}}$ é o valor genético no tempo prévio $j-1$, ρ é o coeficiente de autocorrelação serial e α_j é o efeito genético na nova medição j.

10. Referências Bibliográficas

AKAIKE, H. A new look at the statistical model identification. ***IEEE Trans. Automat.*** *Control*, v. 19, p. 716-723, 1974.

APIOLAZA, L. A.; GILMOUR, A. R.; GRRICK, D. J. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. ***Canadian Journal of Forestry Research***, v. 30, p. 645-654, 2000.

BOX, G. E. P.; JENKINS, G. M. ***Time series analysis*****:** forecasting and control. San Francisco: Holden-Day, 1970. Não paginado.

CARVALHEIRO, R.; FRIES, L. A.; MUNIZ, C. A. S. D.; QUEIRÓZ, S. A.; Estudo de simulação das relações entre a média aritmética, a média harmônica, e o desvio padrão do ganho médio diário de peso do nascimento ao desmame de bovinos. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, p. 38., 2001, Piracicaba. ***Anais***. Piracicaba: FEALQ, 2001. CD ROM.

COSTA e SILVA, J.; DUTKOWSKI, G. W.; GILMOUR, A. R. Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. ***Silvae Genetica***, v. 31, p. 1887-1893, 2001.

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

CULLIS, B. R. ; GOGELL, B.; VERBYLA, A. Spacial analysis of multi-environment early generation variety trials. ***Biometrics***, v. 54, p. 1-18, 1998.

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

DAVIS, J. C. ***Statistics and data analysis in geology*****.** New York: J. Wiley, 1986. 646 p.

DEGEN, B.; SCHOLZ, F. Spatial genetic differentiation among populations of european beech in western germany as identified by geostatistical analysis. ***Forest Genetics***, v. 5, n. 3, p. 191-199, 1998.

DUARTE, J. B. ***Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal***. 2000. 293 f. Tese (Doutorado) - ESALQ/USP, Piracicaba.

EISENBERG, B. E.; GAUCH, H. G.; ZOBEL, R. W.; KILIAN, W. Spatial analysis of field experiments: fertilizer experiments with wheat (*Triticum aestivum*) and tea (*Camellia simensis*). In: KANG, M. S.; GAUCH, H. G. (Ed.). ***Genotype by environment interaction***. Boca Raton: CRC Press, 1996. p. 373-404.

FEDERER, W. T. Recovery of interblock, intergradient and intervarietal information in incomplete block and lattice rectangle designed experiments. ***Biometrics***, v. 54, p. 471-481, 1998.

GALE, J. S.; LAWRENCE, M. J. The decay of variability. In: HOLDEN, J. H. W.; WILLIAMS, J. T. ***Crop genetic resources***: conservation and evaluation. London: Allen and Unwin, 1984. Não paginado.

GILMOUR, A. R.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. ***ASREML reference manual***. Orange: NSW Agriculture, 2000. 218 p.

GILMOUR, A. R.; THOMPSON, R. Moddelling variance parameters in ASREML for repeated measures. In: WORLD CONGRESS ON GENETIC APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. ***Proceedings***. Armidale: AGBU / University of New England, 1998. v. 27, p. 453-454.

GILMOUR, A. R. Post blocking gone too far! Recovery of information and spatial analysis in field experiments. ***Biometrics***, v. 56, p. 944-946, 2000.

GILMOUR, A. R.; CULLIS, B. R.; VERBYLA, A. P. Accounting for natural and extraneous variation in the analysis of field experiments. ***Journal of Agricultural, Biological and Environmental Statistics***, v. 2, p. 269-293, 1997.

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.

GREEN, P. J.; SILVERMAN, B. W. ***Nomparametric regression and generalized linear models***. Chapman & Hall, [S.l.]: 1994. 182p.

GRONDONA, M. O.; CROSSA, J.; FOX, P. N.; PFEIFFER, W. H. Analysis of variety yield trials using two-dimensional separable ARIMA processes. ***Biometrics***, v. 52, p. 763-770, 1996.

KUSNADAR, D.; GALWEY, N. A proposed method for estimation of genetic parameters on forest trees without raising progeny: critical evaluation and refinement. ***Silvae Genetica***, v. 49, p. 15-21, 2000.

LANDIM, P. M. B. ***Análise estatística de dados geológicos***. São Paulo: Fundação Ed. da UNESP, 1998. 226 p.

LECORRE, V.; ROUSSEL, G.; ZANETTO, A.; KREMER, A. Geographical structure of gene diversity in *Quercus petraea*. III. Patterns of variation identified by geostatistical analysis. ***Heredity***, v. 80, p. 464-473, 1998.

MARTIN, R. J. A subclass of lattice processes applied to a problem in planar sampling. ***Biometrika***, v. 66, p. 209-217, 1979.

MARTIN, R. J. On the design of experiments under spatial correlation. ***Biometrika***, v. 73, p. 247-277, 1986.

MARTIN, R. J. Some aspects of experimental design and analysis when errors are correlated. ***Biometrika***, v. 69, p. 597-612, 1982.

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

MARTINEZ, R. Control de la correlación espacial en experimentos de campo en el sector agricola. ***Agronomia Colombiana***, v. 11, p. 83-89, 1994.

MORETTIN, P. A.; TOLOI, C. M. ***Previsão de séries temporais***. 2. ed. São Paulo: Atual, 1987. 439 p.

MUNIZ, C. A. S. D.; CARVALHEIRO, R.; QUEIRÓZ, S. A.; FRIES, L. A. Critérios de seleção baseados em médias aritméticas e médias harmônicas. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 38., 2001, Piracicaba. ***Anais*****.** Piracicaba: FEALQ, 2001. CD ROM.

PANTER, D. M.; ALLEN, F. L. Using best linear unbiased predictions to enhance breeding for yield in soybean. I. Choosing parents. ***Crop Science***, v. 35, p. 379-405, 1995.

PAPADAKIS, J. Advances in the analysis of field experiments. ***Commun. Acad. Athenes***, v. 59, p. 326-342, 1984.

PEARCE, S. C. Field experimentation on rough land: the method of Papadakis reconsidered. ***Journal of Agricultural Science***, v. 131, p. 1-11, 1998.

QIAO, C. G.; BASFORD, K. E.; DELACY, I. H.; COOPER, M. Evaluation of experimental designs and spacial analysis in wheat breeding trials. ***Theoretical and Applied Genetics***, v. 100, p. 9-16, 2000.

RESENDE, M. D. V. de; STURION, J. A. ***Análise genética de dados com dependência espacial e temporal no melhoramento de plantas perenes via modelos geoestatísticos e de séries temporais empregando REML/BLUP individual*.** Colombo: Embrapa Florestas, 2001. 54 p. (Embrapa Florestas. Documentos, 55).

RIBEIRO JUNIOR, P. J. ***Métodos geoestatísticos no estudo da variabilidade espacial de parâmetros do solo***. 1995. 99 f. Dissertação (Mestrado) - ESALQ/ USP, Piracicaba.

SCHOENBERG, I. J. Contributions to the problem of approximation of equidistant data by analitic functions. ***Quaterly Applied Mathematics***, v. 4, p. 44-99, 1946.

SOARES, A. ***Geoestatística para as ciências da terra e do meio ambiente***. Lisboa: IST Press, 2000. 206 p.

STREINER, D. L. Do you see what I mean? Indices of central tendency. ***Canadian Journal of Psychiatry***, v. 45, n. 9, p. 833-836, 2000.

VALENTE, J. M. G. P. ***Geomatemática*:** lições de geoestatística. 2. ed. Ouro Preto: Fundação Gorceix, 1989. v. 3.

VENCOVSKY, R.; CROSSA, J. Variance effective population size under mixed self and random mating with applications to genetic conservation of species. ***Crop Science*,** v. 39, p. 1282-1294, 1999.

VERBYLA, A. P.; CULLIS, B. R.; KENWARD, M. G.; WELHAM, S. J. The analyses of designed experiments and longitudinal data using smootthing splines. ***Journal of the Royal Statistical Society*.** Series C, v. 48, p. 269-311, 1999.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In.: NOVAIS, R. F.; ALVAREZ, V. H.; SCHAEFER, C. E. ***Tópicos em ciência do solo***. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000, p. 1-54.

WALSH, J. L.; AHLBERS, J. H.; NILSON, E. N. Best approximation properties of the spline fit. ***Journal of Mathematical Mecanics***, v. 11, p. 225-234, 1962.

WHITE, I. M. S.; THOMPSON, R.; BROTHERSTONE, S. Genetic and envirommental smoothing of lactation curves with cubic splines. ***Journal of Dairy Science***, v. 82, p. 632-638, 1999.

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.