Bayesian estimation of a surface to account for a spatial trend using penalized splines in an individual-tree mixed model

Eduardo P. Cappa and Rodolfo J.C. Cantet

Abstract: Unaccounted for spatial variability leads to bias in estimating genetic parameters and predicting breeding values from forest genetic trials. Previous attempts to account for large-scale continuous spatial variation employed spatial coordinates in the direction of the rows (or columns). In this research, we use an individual-tree mixed model and the tensor product of B-spline bases with a proper covariance structure for the random knot effects to account for spatial variability. Dispersion parameters were estimated using Bayesian techniques via Gibbs sampling. The procedure is illustrated with data from a progeny trial of *Eucalyptus globulus* subsp. *globulus* Labill. Four different models were used in the sequel. The first model included block effects and the three other models included a surface on a grid of either 8×8, 12×12, or 18×18 knots. The three models with B-splines displayed a sizeable lower value of the deviance information criterion than the model with blocks. Also, the mixed models fitting a surface displayed a consistent reduction in the posterior mean of σ^2_e, an increase in the posterior means of σ^2_A and h^2_{DBH}, and an increase of 66% (for parents) or 60% (for offspring) in the accuracy of breeding values.

Résumé : Le fait de ne pas tenir compte de la variabilité spatiale engendre des biais dans l'estimation des paramètres génétiques et la prédiction des valeurs en croisement qui sont faites à partir des tests en génétique forestière. Les tentatives antérieures pour prendre en compte la variabilité spatiale continue à grande échelle étaient basées sur les coordonnées spatiales dans la direction des rangs (ou colonnes). Dans la présente étude, les auteurs ont employé un modèle mixte basé sur les valeurs d'arbres individuels et le produit tensoriel de B-splines avec une structure de covariance adéquate pour les effets aléatoires de nœuds afin de tenir compte de la variabilité spatiale. Les paramètres de dispersion ont été évalués à l'aide de l'échantillonnage de Gibbs, une méthode basée sur les statistiques bayésiennes. L'approche analytique est illustrée à l'aide de données provenant d'un test de descendances d'*Eucalyptus globulus* subsp. *globulus* Labill. Les auteurs ont utilisé quatre modèles différents dans la suite d'analyses. Le premier modèle incluait des effets de blocs alors que les trois autres modèles impliquaient une surface couchée sur un réseau de 8×8, 12×12 ou 18×18 nœuds. Les trois derniers modèles qui impliquaient des B-splines affichaient une valeur sensiblement plus faible quant au critère d'information sur la déviation comparativement au modèle avec les blocs. De plus, les modèles mixtes qui s'ajustaient à une surface démontraient une réduction constante de leur moyenne a posteriori de σ^2_e, une augmentation de leurs moyennes a posteriori de σ^2_A et de h^2_{DBH}, et une augmentation de 66 % (pour les parents) et de 60 % (pour les descendants) de la précision des valeurs en croisement.

[Traduit par la Rédaction]

Introduction

Forest genetic trials are prone to a high degree of environmental heterogeneity as compared with other cultivated plants (Libby and Cockerham 1980): trees are large living organisms and occupy more space than most cultivated plant species. Moreover, trees are often planted in places with heterogeneous levels of fertility, humidity, soil depth, or slope. Although spatial heterogeneity is a nuisance effect in forest genetic evaluation where the main goal is the prediction of breeding values, ignoring such a source can lead to biases in the estimation of genetic parameters and the prediction of individual additive genetic effects (breeding values: Magnussen 1993, 1994). To account for environmental gradients, tree breeders have devised forest trials using randomized complete blocks or incomplete block designs. However, setting fixed limits for the blocks makes it difficult to account for continuously varying environmental factors. Additionally, establishing a priori a design that properly accounts for all sources of environmental heterogeneity may be a hopeless task as "environmental variation is never known prior to establishment" (Fu et al. 1999*a*). Alternatively, the spatial variation can be accounted for a posteriori within the model of evaluation. In these so called "spatial models", the continuous variability has two main sources: the "local trend", or small-scale variation, and the "global trend", or large-

Received 17 October 2006. Accepted 18 June 2007. Published on the NRC Research Press Web site at cjfr.nrc.ca on 9 January 2008.

E.P. Cappa.[1,2] Fondo para la Investigación Científica y Tecnológica (FONCyT), Argentina.
R.J.C. Cantet. Department of Animal Production, University of Buenos Aires, Avenida San Martín 4453, C1417DSQ Buenos Aires, Argentina; Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

[1]Corresponding author (e-mail: ecappa@mail.agro.uba.ar).
[2]Present address: Department of Animal Production, University of Buenos Aires, Avenida San Martín 4453, C1417DSQ Buenos Aires, Argentina.

 doi:10.1139/X07-116 © 2007 NRC Canada

scale variation, across a spatial gradient. The two sources are observable in forest genetic trials: either component alone or in combination with each other (e.g., Fu et al. 1999*b*; Costa e Silva et al. 2001; Dutkowski et al. 2002).

Models that account for large-scale continuous spatial variation include spatial coordinates expressed as either classification variables or covariables. The latter are nonstochastic functions such as polynomials (Federer 1998) or smoothing splines (Verbyla et al. 1999). Costa e Silva et al. (2001) and Dutkowski et al. (2002) considered the continuous spatial trends by fitting a Kronecker product of first-order autoregressive (AR(1)) covariance structures for rows and columns (Gilmour et al. 1997). To account for large-scale variation, Costa e Silva et al. (2001) proposed the use of fixed or random classification variables, and Dutkowski et al. (2002) included fixed effects of spatial coordinates as quadratic polynomials or cubic smoothing splines (Verbyla et al. 1999). However, Dutkowski et al. (2002) found that the variograms in two out of five trials were not stationary, implying that large-scale covariance was still present in the spatial errors. Notice that fitting classification variables, either in the additive model with row and column effects or in the model with interaction, does not treat the information in a continuous fashion so that a surface cannot be fitted. On the other hand, additive models with polynomials or splines use only marginalized estimates of row and columns effects so that the information on the inner points of the grid is lost. Therefore, in forest genetic trials where continuous spatial variation develops in two dimensions, analyses using classification variables, covariables, polynomials, or splines in one dimension, or additive models, or a model with interaction of classification variables may not completely account for spatial covariance. Thomson and El-Kassaby (1988) fitted sixth-order-degree polynomials in two dimensions by least-squares to compare different provenances of Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco). The use of polynomials for the analysis in two dimensions ("trend analysis") of forest genetic data can also be found in the work of Liu and Burkhart (1994) and Saenz-Romero et al. (2001). However, the fit of polynomials suffers from several drawbacks (Green and Silverman 1994, p. 2). First of all, the fit is global and not local, which means that (*i*) the method is not capable of accounting for local variations present in the data, (*ii*) few influential observations may exert a large influence in the resulting fit, and (*iii*) the fit in the edges is usually poor. Another serious drawback with polynomials is their numerical instability as the order of the polynomial increases.

Splines are a more efficient approach to the use of polynomials. They are segmented polynomial functions that are locally fitted such that the resulting function is differentiable at the joints of the segments ("knots") up to the order of fit. Splines are able to capture most sinuosities present in the data and do not suffer from numerical instability. Eilers and Marx (1996) introduced "penalized splines" (P-splines) in one dimension using basic splines (B-splines) with equally spaced knots and a linear model approach with a roughness penalty consisting of the differences among the parameters, i.e., the effects of the knots. T. Speed (see Robinson 1991) first pointed out the connection between splines and mixed models, a subject further expanded by Ruppert et al. (2003) and Wand (2003). Cantet et al. (2005) approached P-splines in one dimension using proper covariance structures rather than matrices of differences in an animal breeding context. Eilers and Marx (2003) extended their methodology to estimate a surface along two dimensions using the tensor product of B-splines. Either in one or in two dimensions (Eilers and Marx 1996, 2003), the parameters of the B-spline function are treated as fixed effects. Similar results can be obtained by a mixed model approach by treating the B-spline function parameters as random variables, which from now on we refer to as "random knot effects" (RKE). The goal of the present research is to show how to fit a surface using the tensor product of B-spline bases to account for large-scale continuous spatial variation in an individual-tree mixed model for forest genetic evaluation. To do that, we superimpose a covariance structure for the RKE in a two-dimensional grid. As in some recent contributions to forest breeding (e.g., Soria et al. 1998; Cappa and Cantet 2006; Waldmann and Ericsson 2006), we employed the Bayesian approach via Gibbs sampling to make inferences in all dispersion parameters of the model. Developments are illustrated by means of data on diameter at breast height from a progeny trial of *Eucalyptus globulus* subsp. *globulus* Labill. The resulting estimates of all dispersion parameters for mixed models that include the fitted surface are finally compared with corresponding estimates from the classical model including blocks.

Methods

Two-dimensional tensor product of B-splines

We first briefly introduce P-splines in one dimension as suggested by Eilers and Marx (1996). Then, we take the approach of Eilers and Marx (2003) and Green and Silverman (1994) and extend P-splines to two dimensions using the tensorial product of B-splines.

Eilers and Marx (1996) advocated using B-splines with equally spaced knots to obtain P-splines. B-splines are local basis functions consisting of polynomial segments of degree d, in general quadratic or cubic, that have $d - 1$ continuous derivatives at the joining points, or knots. Therefore, knots are the parametric values where the polynomial functions that compose splines join one another. We will denote with nx the number of knots of the spline function. A B-spline of degree d is positive on a domain spanned by $d + 2$ knots and is zero elsewhere. All in all, $d + 1$ B-spline coefficients are nonzero. Eilers and Marx (1996) introduced a penalty that affects first or second differences of B-spline parameters. The penalty controls the degree of smoothness while fitting the function. Let the data vector of length n be denoted by $\boldsymbol{y}$. Also, let the vector $\boldsymbol{x}$ contain the positions of the rows (or columns) of trees expressed as a distance (in metres) from the beginning row (or column) of the trial. For example, with a spacing of 3 m between rows, $\boldsymbol{x}' = [0, 3, 6, 9, ...]$. Then, the one-dimensional spline function $\boldsymbol{s}(\boldsymbol{x})$ for rows (or columns) is written as

$$[1] \qquad \boldsymbol{s}(\boldsymbol{x}) = \sum_{i=1}^{nx} \boldsymbol{B}_i(x)\boldsymbol{b}_i$$

where $\boldsymbol{B}_i = (\boldsymbol{B}_1(x), \boldsymbol{B}_2(x), ..., \boldsymbol{B}_{nx}(x))'$ is a column vector with nx B-spline bases (De Boor 1993) and $\boldsymbol{b}_i = (\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_{nx})'$

© 2007 NRC Canada

denotes the vector of spline parameters in one dimension. To express one row (or column) in $\boldsymbol{x}$ as a function of B-spline bases, two linear B-splines bases are needed, or three quadratic B-splines bases, or four cubic B-splines bases. In matrix form, expression 1 can be written as $\mathbf{B}\boldsymbol{b}$, where $\mathbf{B}$ is the $n \times nx$ matrix that contains the $\boldsymbol{B}_i$s, and $\boldsymbol{b}$ is the parametric vector ($nx \times 1$) containing the $\boldsymbol{b}_i$s to form $\boldsymbol{s}(\boldsymbol{x})$. The functional 1 is generally fitted by least-squares with an additive penalty. Eilers and Marx (1996) observed that the penalized estimator of $\boldsymbol{b}$ is the solution of the following system of equations:

$$[2] \qquad (\mathbf{B}'\mathbf{B} + \lambda \mathbf{D}_d'\mathbf{D}_d)\widehat{\boldsymbol{b}} = \mathbf{B}'\boldsymbol{y}$$

where the positive scalar λ controls the amount of smoothing and $\mathbf{D}_d$ is the matrix of differences of order d. For $d = 1$ and $d = 2$, we have, respectively

$$[3] \qquad \mathbf{D}_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$\mathbf{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

Ruppert et al. (2003) and Wand (2003) discussed the connection between P-splines and mixed models (Henderson 1984). The smoothing parameter λ is seen as the ratio of the error variance to the variance of the B-spline parameters $\boldsymbol{b}_i$. Moreover, $\mathbf{D}'\mathbf{D}$ is interpreted as a g-inverse of the covariance matrix of the B-spline RKE (Cantet et al. 2005) and acts as a singular penalization matrix.

Tensor products of B-splines allow a natural extension of one-dimensional P-spline smoothing to two dimensions by means of the Kronecker product of single structures. A more rigorous approach can be found in Green and Silverman (1994, pp. 155–159). The tensor product of two univariate B-splines along the rows (r) and columns (c) is defined as the $r \times c$ rectangle in $\Re^2$ such that $\boldsymbol{T}_{kl}(r,c) = \boldsymbol{Br}_k(r)\boldsymbol{Bc}_l(c)$, where $\boldsymbol{Br}_k(r)$ and $\boldsymbol{Bc}_l(c)$ are B-spline bases for the row ($k = 1, 2, ..., nx_r$) and column ($l = 1, 2, ..., nx_c$) RKE, respectively. The scalar nx_r is the number of knots for rows and nx_c is the number of knots for columns. If row and column knots are chosen to be equally spaced and cubic B-splines are used, the $r \times c$ space can be divided into small rectangular panels such that $[r_k, r_{k+1}] \times [c_k, c_{k+1}]$. Let $\mathbf{S} = [\gamma_{kl}]$ be the $nx_r \times nx_c$ matrix containing the parameters from the tensor product of B-splines that have to be estimated. Then, for a given set of knots, the surface $\omega(r,c)$ can be approximated using the following matrix expression:

$$[4] \qquad \mathrm{vec}\{\omega(r,c)\} = \mathbf{B}\boldsymbol{b}$$

where $\mathbf{B}$ has dimension $n \times (nx_r \times nx_c)$ and is equal to $\mathbf{B} = (\mathbf{B}_r \otimes \mathbf{1}'_{nx_c})\#(\mathbf{1}'_{nx_r} \otimes \mathbf{B}_c)$. The notation "vec" stands for the operator that results from stacking the columns of a matrix into a vector (Harville 1997, p. 339), and the symbols $\otimes$ and # indicate the Kronecker and Hadamard products of matrices, respectively (Harville 1997). The order of the resulting vector $\boldsymbol{b}$ is $(nx_r \times nx_c) \times 1$.

In analogy to what they had done for one dimension (Eilers and Marx 1996), Eilers and Marx (2003) and Marx and Eilers (2005) proposed a two-dimensional penalized estimation of a surface. Let λ_r and λ_c be the parameters controlling the degree of smoothness for rows and columns, respectively, whereas $\mathbf{D}_r$ and $\mathbf{D}_c$ are the respective difference matrices 3. Then, the solution for $\widehat{\boldsymbol{b}}$ is obtained by solving the equations

$$[5] \qquad \left(\mathbf{B}'\mathbf{B} + \lambda_r(\boldsymbol{I}_{nx_r} \otimes \mathbf{D}_r'\mathbf{D}_r) + \lambda_c\ (\mathbf{D}_c'\mathbf{D}_c \otimes \boldsymbol{I}_{nx_c})\right)\widehat{\boldsymbol{b}} = \mathbf{B}'\boldsymbol{y}$$

The expression above is similar to the system in one dimension where $\mathbf{B}$ is replaced by $\boldsymbol{Br}$ or $\boldsymbol{Bc}$ and $\lambda\mathbf{D}'\mathbf{D}$ is replaced by $\lambda_r\ (\boldsymbol{I}_{nx_r} \otimes \mathbf{D}_r'\mathbf{D}_r) + \lambda_c\ (\ \mathbf{D}_c'\mathbf{D}_c \otimes \boldsymbol{I}_{nx_c})$. In the next section, we show how to fit data in two dimensions using the tensor product of B-splines by means of a mixed linear model.

Mixed model representation of a two-dimensional tensor product of B-splines

In forest genetic trials, trees are usually arranged in regular grids arrayed in rows and columns. To position any tree, let r and c be the row and column coordinates, respectively, measured in metres or degrees. Let $\mathbf{Y}$ be a matrix of order n_r (number of rows) $\times$ n_c (number of columns) containing the observations for a trait (such as height or diameter). To transform $\mathbf{Y}$ into a vector, we again use the vec operator in which the $n_r n_c \times 1$ vector $\boldsymbol{y}$ results from stacking the columns of $\mathbf{Y}$:$\boldsymbol{y}$ = vec($\mathbf{Y}$). Then, an individual-tree mixed model with a smoothed surface to account for environmental heterogeneity is given by

$$[6] \qquad \boldsymbol{y} = \mathbf{X}\beta + \mathbf{B}\boldsymbol{b} + \mathbf{Z}\boldsymbol{a} + \boldsymbol{e}$$

where β is a $p \times 1$ vector of fixed effects associated with $\boldsymbol{y}$ by the incidence matrix $\mathbf{X}(n \times p)$ such that the matrix $\mathbf{X}$ is of full column rank p. In case the rank of $\mathbf{X}$ is less than p, it is always possible to find a reparametrization that turns $\mathbf{X}$ into a matrix of full-column rank (Christensen 1987). The random $q \times 1$ vector $\boldsymbol{a}$ contains the additive genetics effects (or breeding values) of individual trees and is related to $\boldsymbol{y}$ by the incidence matrix $\mathbf{Z}$ (of order $n \times q$). The expectation of $\boldsymbol{a}$ is $\mathbf{0}$ and the covariance matrix is $\mathbf{A}\sigma_A^2$ where $\mathbf{A}$ is the additive tree-level relationship matrix (Henderson 1984) for the trial trees and their known ancestors and σ_A^2 is the additive genetic variance. The distribution of the random ($nx_r \times nx_c$) $\times$ 1 vector $\boldsymbol{b}$ containing the parameters of the tensor product of B-splines (i.e., the RKE) is such that $\boldsymbol{b} \sim N(\mathbf{0}, \mathbf{U}\sigma_b^2)$. The scalar σ_b^2 is the variance of the RKE for rows and columns and $\mathbf{U}$ is the covariance structure in two dimensions. Finally, random error terms are included in the $n \times 1$ vector $\boldsymbol{e}$, which is distributed as $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ and σ_e^2 is the error variance.

The covariance structure $\mathbf{U}$ plays an important role in model 6. The matrix should reflect the correlation decay among B-spline knots that are farther apart, either row- or column-wise. A possible choice for $\mathbf{U}$ is $\boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c$, a Kronecker product of matrices for the rows ($\boldsymbol{\Sigma}_r$) and for the columns ($\boldsymbol{\Sigma}_c$). If $\mathbf{U}$ is a linear covariance structure (Anderson 1973), the estimation process is simplified and there is only one parameter to estimate: σ_b^2. Then, estimation can be performed with simpler methods and algorithms, i.e., REML-EM

© 2007 NRC Canada

or Gibbs sampling. The challenge is to find a $\mathbf{U}$ that is informative enough among the correlation decay among knot effects and at the same time that does not depend on extra parameters. A possibility is to set $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_c$ to be equal to the one-dimensional covariance structure originally proposed by Green and Silverman (1994, p. 13) and then used by Durban et al. (2001) to fit a fertility trend. In this tridiagonal matrix, correlations are nonzero for neighbor knots and are zero otherwise. More explicitly, if ζ_{ij} is element ij of any of the matrices $\mathbf{\Sigma}_r$ or $\mathbf{\Sigma}_c$, diagonals are ζ_{ii} = 4/6, whereas off-diagonals are either $\zeta_{i+1,i} = \zeta_{i,i+1}$ = 1/6 or ζ_{ij} = 0 for $|i - j| \geq 2$, $i = j = 1, 2, ..., nx_r$ or nx_c. Thus, besides being positive definite, $\mathbf{U} = \mathbf{\Sigma}_r \otimes \mathbf{\Sigma}_c$ is strictly diagonally dominant as $|\varsigma_{ii}| > \sum_{j \neq i} |\varsigma_{ij}|$ for every i. To exemplify, suppose $nx_r = nx_c = 4$, then

$$\mathbf{\Sigma}_r = \mathbf{\Sigma}_c = \frac{1}{6}\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

and $\mathbf{U} = \mathbf{\Sigma}_r \otimes \mathbf{\Sigma}_c$ is given by

$$\frac{1}{6}\begin{bmatrix}
16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 4 & 0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 & 4 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 & 1 & 4 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 0 & 0 & 4 & 16 & 0 & 0 & 1 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 1 & 0 & 0 & 16 & 4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 1 & 0 & 4 & 16 & 4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 0 & 0 & 4 & 16
\end{bmatrix}$$

In this example, nonzero elements of $\mathbf{U}$ are correlations between neighbor knots. Take, for example, the second knot (row 2 of $\mathbf{U}$) having as proximal neighbors the knots 1, 3, and 6 and as diagonal neighbors the knots 5 and 7. Notice that correlations with neighbors in proximal positions are stronger (4/6) than with neighbors located diagonally (1/3). Implicit is the assumption that the spacing between both columns and rows is equal. There are other structures that allow modeling a gradual decay in correlation as knots are separated farther in the direction of the rows or of the columns, such as those proposed by Hyndman et al. (2005) or Cantet et al. (2005). Finally, given the random effects in eq. 6, the covariance matrix $\mathbf{y}$ (say $\mathbf{V}$) is as follows:

$$[7] \qquad \mathbf{V} = \mathbf{ZAZ}'\sigma_A^2 + \mathbf{BUB}'\sigma_b^2 + \mathbf{I}_n\sigma_e^2$$

and mixed model equations (Henderson 1984) for eq. 6 are

$$[8] \qquad \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B} & \mathbf{X}'\mathbf{Z} \\ \mathbf{B}'\mathbf{X} & \mathbf{B}'\mathbf{B} + \mathbf{U}^{-1}\lambda & \mathbf{B}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{B} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \times \begin{bmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{y} \\ \mathbf{B}'\boldsymbol{y} \\ \mathbf{Z}'\boldsymbol{y} \end{bmatrix}$$

where $\lambda = \sigma_e^2/\sigma_b^2$ and $\alpha = \sigma_e^2/\sigma_A^2$. Notice that in the Bayesian view of the mixed linear model (Sorensen and Gianola 2002), the likelihood of the data is proportional to

$$[9] \qquad p(\boldsymbol{y}|\beta, \boldsymbol{a}, \boldsymbol{b}) \propto (\sigma_e^2)^{-\frac{1}{2}} \times \exp \times \left[-\frac{1}{2\sigma_e^2}(\boldsymbol{y} - \mathbf{X}\beta - \mathbf{Za} - \mathbf{Bb})'(\boldsymbol{y} - \mathbf{X}\beta - \mathbf{Za} - \mathbf{Bb})\right]$$

Bayesian estimation

The Bayesian approach via Gibbs sampling was used to estimate the parameters in model 6 (Sorensen and Gianola 2002). We now specify the prior distributions as well as the joint and marginal conditional posterior densities.

Specification of prior distributions

Conjugate prior densities were chosen for all parameters. To reflect a prior state of uncertainty for the fixed effects and to keep a proper posterior distribution (Hobert and Casella 1996), we set $\beta \sim N_p(0, \mathbf{K})$ and $\mathbf{K}$ is a diagonal matrix with very large elements ($k_{ii} > 10^8$). Therefore, this prior density is proportional to

$$[10] \qquad p(\beta|\mathbf{K}) \propto |\prod_{i=1}^{p} k_{ii}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\frac{\beta_i^2}{k_{ii}}\right\}$$

The Bayesian view of mixed model 6 is that of a hierarchical model (e.g., see Hobert and Casella 1996, expression 5,

© 2007 NRC Canada

p. 1462). Under this approach, the vectors $\boldsymbol{b}$ and $\boldsymbol{a}$ are considered first stage parameters, whereas the dispersion parameters of the distributions of $\boldsymbol{b}$ and $\boldsymbol{a}$ are second-stage parameters, such as the variance components σ^2_b and σ^2_A. Hence, we define the prior distribution of $\boldsymbol{b}$ conditional on σ^2_b and the prior distribution of $\boldsymbol{a}$ conditional on σ^2_A. Then, the vector $\boldsymbol{b}$ is distributed a priori as $\boldsymbol{b} \sim N_b(0, \mathbf{U}\sigma^2_b)$ so that

$$[11] \qquad p\,(\boldsymbol{b}|\sigma^2_b) \propto (\sigma^2_b)^{-\frac{nx\times nx}{2}} \exp\left\{ -\frac{\boldsymbol{b}'\mathbf{U}^{-1}\boldsymbol{b}}{2\sigma^2_b} \right\}$$

The prior density for the vector of breeding values is $\boldsymbol{a} \sim N_q(0, \mathbf{A}\sigma^2_A)$ (see eq. 13.2 in Sorensen and Gianola 2002, p. 564) so that

$$[12] \qquad p\,(\boldsymbol{a}|\sigma^2_A) \propto (\sigma^2_A)^{-\frac{q}{2}} \exp\left\{ -\frac{\boldsymbol{a}'\mathbf{A}^{-1}\boldsymbol{a}}{2\sigma^2_A} \right\}$$

Following Sorensen and Gianola (2002), we chose to use independent scaled inverted χ^2 densities as prior distributions for the variance components σ^2_b, σ^2_A. and σ^2_e:

$$[13] \qquad p(\sigma^2_b|\upsilon_b, \delta^2_b) \propto (\sigma^2_b)^{-\left(\frac{\upsilon_b}{2}+1\right)} \exp\left\{ -\frac{\upsilon_b\delta^2_b}{2\sigma^2_b} \right\}$$

$$[14] \qquad p(\sigma^2_A|\upsilon_A, \delta^2_A) \propto (\sigma^2_A)^{-\left(\frac{\upsilon_A}{2}+1\right)} \exp\left\{ -\frac{\upsilon_A\delta^2_A}{2\sigma^2_A} \right\}$$

$$[15] \qquad p(\sigma^2_e|\upsilon_e, \delta^2_e) \propto (\sigma^2_e)^{-\left(\frac{\upsilon_e}{2}+1\right)} \exp\left\{ -\frac{\upsilon_e\delta^2_e}{2\sigma^2_e} \right\}$$

Parameters in the densities 13, 14, and 15 are the hypervariances δ^2_b, δ^2_A, and δ^2_e and the degrees of freedom υ_b, υ_A, and υ_e, respectively.

Joint and conditional posterior densities

By multiplying eq. 9 by eqs. 10, 11, 12, 13, 14, and 15, the joint posterior density of all parameters is proportional to

$$[16] \qquad p(\beta, \boldsymbol{a}, \boldsymbol{b}, \sigma^2_b, \sigma^2_A, \sigma^2_e|\boldsymbol{y}, \upsilon_b, \upsilon_A, \upsilon_e, \delta^2_b, \delta^2_A, \delta^2_e) \propto p(\boldsymbol{y}|\beta, \boldsymbol{a}, \boldsymbol{b})p(\beta|\mathbf{K})p(\boldsymbol{b}|\sigma^2_b)p(\boldsymbol{a}|\sigma^2_A)p(\sigma^2_b|\upsilon_b\delta^2_b)p(\sigma^2_A|\upsilon_A\delta^2_A)p(\sigma^2_e|\upsilon_e\delta^2_e)$$

Inference on any parameter by means of the Gibbs sampler requires conditional posterior densities in close form. The joint conditional density of β, $\boldsymbol{b}$, and $\boldsymbol{a}$ is given by

$$[17] \qquad \begin{bmatrix} \beta \\ \boldsymbol{b} \\ \boldsymbol{a} \end{bmatrix} \boldsymbol{y}, \sigma^2_A, \sigma^2_b, \sigma^2_e \sim N\left(\begin{bmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{bmatrix}, \begin{bmatrix} \mathbf{X}'\mathbf{X}+\mathbf{K}^{-1} & \mathbf{X}'\mathbf{B} & \mathbf{X}'\mathbf{Z} \\ \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B}+\mathbf{U}^{-1}\lambda & \mathbf{X}'\mathbf{Z} \\ \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B} & \mathbf{X}'\mathbf{Z}+\mathbf{A}^{-1}\alpha \end{bmatrix}^{-1} \right)$$

Vectors $\widehat{\beta}$, $\widehat{\boldsymbol{b}}$, and $\widehat{\boldsymbol{a}}$ are the solutions to the following set of equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X}+\mathbf{K}^{-1} & \mathbf{X}'\mathbf{B} & \mathbf{X}'\mathbf{Z} \\ \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B}+\mathbf{U}^{-1}\lambda & \mathbf{X}'\mathbf{Z} \\ \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B} & \mathbf{X}'\mathbf{Z}+\mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{y} \\ \mathbf{X}'\boldsymbol{y} \\ \mathbf{X}'\boldsymbol{y} \end{bmatrix}$$

We sample the elements of the vectors β, $\boldsymbol{b}$, and $\boldsymbol{a}$ on a one-by-one basis using the procedure discussed by Sorensen and Gianola (2002, expressions 13.11 and 13.12, p. 566). The conditional posterior distribution of σ^2_A is scaled inverted χ^2:

$$[18] \qquad p(\sigma^2_A|\beta, \boldsymbol{b}, \boldsymbol{a}, \sigma^2_b, \sigma^2_e, \boldsymbol{y}) \propto \text{Inv} - \chi^2(\tilde{\upsilon}_A, \tilde{\delta}^2_A)$$

with parameters $\tilde{\upsilon}_A = q + \upsilon_A$ and $\tilde{\delta}^2_A = (\boldsymbol{a}'\mathbf{A}^{-1}\boldsymbol{a} + \upsilon_A\delta^2_A)/\tilde{\upsilon}_A$. Also, for σ^2_b, we have

$$[19] \qquad p(\sigma^2_b|\beta, \boldsymbol{b}, \boldsymbol{a}, \sigma^2_A, \sigma^2_e, \boldsymbol{y}) \propto \text{Inv} - \chi^2(\tilde{\upsilon}_b, \tilde{\delta}^2_b)$$

with $\tilde{\upsilon}_b = nx \times nx + \upsilon_b$ and $\tilde{\delta}^2_b = (\boldsymbol{b}'\mathbf{U}^{-1}\boldsymbol{b} + \upsilon_b\delta^2_b)/\tilde{\upsilon}_b$. Finally, the error variance has the following conditional posterior:

$$[20] \qquad p\left(\sigma^2_e|\beta, \boldsymbol{a}, \boldsymbol{b}, \sigma^2_b, \sigma^2_A, \boldsymbol{y}\right) \propto (\sigma^2_e)^{-\left(\frac{n+\upsilon_e+2}{2}+1\right)} \times \exp\left\{ -\frac{\tilde{\upsilon}_e\tilde{\delta}^2_e}{2\sigma^2_e} \right\}$$

with $\tilde{\upsilon}_e = n + \upsilon_e$ degrees of freedom and scale parameter $\tilde{\delta}^2_e = (\boldsymbol{e}'\boldsymbol{e} + \upsilon_e\delta^2_e)/\tilde{\upsilon}_e$. At any iteration of the Gibbs algorithm, we first sampled from distribution 17, then from eq. 20, then from eq. 18, and finally from eq. 19 to start the process back again. A program was written in FORTRAN to perform all calculations (the FORTRAN program is available from the first author on request).

A working example: analysis of an *E. globulus* subsp. *globulus* progeny trial

Data

An *E. globulus* subsp. *globulus* progeny trial was used in the study. The trial was chosen, as there was indication of spatial trends affecting the data, making the information suitable to illustrate the methodology. The data were collected at Licenciado Matienzo (37°59′578″S, 59°00′107″W) in the southeastern part of Buenos Aires province, Argentina, where *E. globulus* has traditionally being planted (Lopez et al. 2001). The soil was a Petrocalcic Paleudoll. It is characterized by a fine texture and a subsuperficial petrocalcic horizon starting at 0.81 m from surface and with vari-

© 2007 NRC Canada

able depth, and soil capability is III. The A horizon is made of loam and has a granular structure, pH = 6, and 6.2% organic matter. The textures of the B and BC horizons are clay and silty clay loam. The trait was diameter at breast height (DBH, 1.3 m) measured in centimetres when trees were 6 years old on all surviving trees. Records from stunted and damaged trees were removed from the analysis. The resulting data set consisted of 1021 recorded trees from 72 seed lots: 36 open-pollinated families from eight native stand sites in Australia, 28 open-pollinated families, and eight bulk collections from land race originated in Argentina, Portugal, Spain, and Chile. A detailed description of the genetics materials used in this study can be found in Lopez et al. (2001, table 1, p. 245). Fourteen genetic groups were formed according to provenance. After including all known genetic relationships, a total of 1089 individual trees (1021 plants with data plus 68 ancestors) were used in the pedigree file. In calculating the relationship matrix, it was assumed that both the selfing rate and the inbreeding were zero. There were only 113 bulk trees (assumed to be unrelated) in the data set so that no extra independent additive variance was assumed for these individuals. Trees were planted in single-tree plots on a rectangular grid of 32 rows and 36 columns (93 m × 105 m) arrayed in squares of 3 m × 3 m, with 15 replicates per family. Then, rows have coordinates r_i, $i = 1, 2, ..., R = 32$, and columns coordinates c_j, $j = 1, 2, ..., C = 36$. To fit the surface and for numerical purpose, row (r) and column (c) spatial coordinates were expressed in metres and the origin was taken to be the north corner. The first tree ($r = 1$, $c = 1$) was set to coordinates (0, 0) so that $R = 93$ m and $C = 105$ m. As a continuous scale (in metres) is used to fit the data, the approach can be used for planting designs when the distance between rows and columns is not the same.

Models of analysis

Four individual additive tree models were evaluated. All models included a fixed effect of genetic group to account for the means of the different origins of parents, random additive genetic effects (breeding values), and random errors. Model 1 also included fixed block effects. In the other three models (2, 3, and 4), a surface was fitted using the tensor products of cubic B-splines. These models differ in the number of knots: 8 × 8, 12 × 12, and 18 × 18 for models 2, 3, and 4, respectively. The minimum number of knots was chosen using the criterion suggested by M. Wand (see Ruppert 2002) who chose to place a knot every t observations and $t = \min(R/4 \text{ (or } C/4), 35)$. As in P-spline methodology, knots are equally spaced; once the number of knots is chosen, the issue of the placement of the knots does not arise (Eilers and Marx 1996). The coefficients for the cubic B-splines in **B** were calculated using the recursive algorithm of De Boor (1993), and the order of the resulting matrix was $n \times (nx_r \times nx_c)$. Accordingly, the vector $\boldsymbol{b}$ was of order $(nx_r \times nx_c) \times 1$ and the covariance structure **U** of order $(nx_r \times nx_c) \times (nx_r \times nx_c)$. In these data, $n = 1021$ and $nx_r = nx_c = 8$, 12, and 18. The deviance information criterion (DIC) (Spiegelhalter et al. 2002) was employed to compare the fit from different models. The model with the smallest value of DIC should be favored, as this indicates a better fit and a lower degree of model complexity. Numerical details for the calculus of DIC in individual tree models are given in Cappa and Cantet (2006).

To select the covariance matrix of the RKE, we tried the covariance structures used by Cantet et al. (2005), Hyndman et al. (2005), and Durban et al. (2001). The three different matrices **U** were fitted to the model with 12 × 12 knots, and model comparison was based on the DIC. Although the estimates of σ^2_A and σ^2_e from the models with the three covariance structures were similar, the values of the DIC were 2860.87, 2851.19, and 2833.46 for the **U** matrices used by Cantet et al. (2005), Hyndman et al. (2005), and Durban et al. (2001), respectively, which supports the use of the latter structure to analyze the current data set.

Further model comparison was provided by the accuracy of prediction of breeding values, which was computed using the following expression:

$$r = \sqrt{\frac{1 - \text{PEV}}{\sigma^2_A}}$$

The acronym PEV stands for "prediction error variance" (Henderson 1984) of predicted breeding values using the "best linear unbiased predictors" (BLUPs) of parent and offspring. The PEV is calculated as the diagonal elements of the inverse of the coefficient matrix from the mixed model equations (Henderson 1984) in eq. 8. The required variance components to set up the mixed model equations were those estimated from the Bayesian analysis. Spearman correlations were also calculated to compare whether the ranking of predicted breeding values differed among models.

Spatial analysis of residuals

To identify spatial patterns in the data, we examined the spatial distribution as in Gilmour et al. (1997) using a model with fixed genetic groups and random breeding values. The distribution of the DBH residuals is displayed in Fig. 1. The color intensity represents the magnitude of the residuals: the darker the dot, the larger the residual value. Additionally, residuals were plotted against row and column position to detect dissimilar patterns in any row (across columns, Fig. 2*a*) or in any column (across rows, Fig. 2*b*). To exemplify, only rows 1, 16, and 32 and columns 1, 16, and 32 are displayed. Notice the different residual patterns across rows or columns, which indicate the presence of interaction between row and column position and the need for a two-dimensional smoothing.

Computational details and posterior inference

The values of the hypervariances δ^2_A and δ^2_e were estimated from the same data set using an empirical Bayes approach via Gibbs sampling, with an individual-tree model including fixed effects of blocks and genetic groups and random additive genetic effects. As there was no prior information on the hypervariance of the RKE, we tried different values of δ^2_b in the interval $[0, \delta^2_e)$ and found that the algorithm converged always to the same posterior mean of σ^2_b in all spatial models. The degrees of belief were set to 10 (i.e., $n_A = \nu_k = 10$) to reflect a relatively high degree of uncertainty. The DIC was computed for

© 2007 NRC Canada

each model using the output from the Gibbs sampling. At the end of each iteration, heritability of DBH was calculated as $h^2_{DBH} = \tilde{\sigma}^2_A/(\tilde{\sigma}^2_A + \tilde{\sigma}^2_e)$, where $\tilde{\sigma}^2_A$ and $\tilde{\sigma}^2_e$ are the values of the additive and error variance sampled at a given iteration.

A single Gibbs chain of 1 010 000 iterations was sampled, and the first 10 000 iterates were discarded due to "burn-in". This number was calculated using the diagnostics of Raftery and Lewis (1992) as implemented in "Bayesian output analysis" (BOA) version 1.0 (Smith 2003). This software was also used to calculate the autocorrelations for all lags from 1 to 50. To evaluate the impact of autocorrelations in the variability of the samples, the "effective sample size" proposed by R. Neal (Kass et al. 1998) was calculated for each parameter as

$$\mathrm{ESS} = \frac{1\ 000\ 000}{1 + 2\sum_{i=1}^{50} \rho(i)}$$

where $\rho(i)$ is the autocorrelation measured at lag i. Marginal posterior densities for the variance components were estimated by the Gaussian kernel method (Silverman 1986, chap. 2):

$$[21] \qquad f(\theta_i) = \frac{1}{1\ 000\ 000\ h} \sum_{j=1}^{1\,000\,000} \frac{1}{\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left(\frac{z-\theta_{ij}}{h}\right)^2\right]$$

where $f(\theta_i)$ is the estimated posterior density of θ_i ($i = \sigma^2_A$, σ^2_b, or σ^2_e) and θ_{ij} is the jth (j = 1, ..., 1 000 000) sampled value of variance component θ_i. The scalar h is the window width estimated by unbiased cross-validation. Mean, mode, median, standard deviation, and 95% high posterior density interval were then calculated with BOA for all parameters from the individual marginal posteriors using the free software R (www.r-project.org/).

Results

The values of DIC for models 1–4 were 3152.66, 2868.64, 2833.46, and 2835.12, respectively. Note that all models including a tensor product of B-splines had a smaller DIC (i.e., better fits) than model 1 with block effects. Model 3 (12 × 12 knots) showed the smallest DIC, closely followed by model 4 (18 × 18 knots). The presence of spatial effects can be observed in Fig. 3, which displays the estimates of the block effects for model 1 or the estimated surface for models 2–4. There seems to be similarities in the locations of the high and low areas in all four contour plots. The fit for model 1 is expectedly abrupt, as block effects are parameters for a categorical variable. On the other hand, the estimated surfaces with models 2–4 show that the degree of smoothness decreases with the increase in the number of knots from 8 to 18.

Posterior statistics for σ^2_A, σ^2_b, σ^2_e, and h^2_{DBH} are shown in Table 1. Posterior means, medians, and modes of the variance components and h^2_{DBH} were similar except for σ^2_A from models 2 and 3 and σ^2_e from model 1, where the modes were smaller than the means and medians. Estimates of σ^2_A and σ^2_e were similar in models 2–4, and this resulted in similar posteriors means of h^2_{DBH}: 0.243, 0.267, and 0.262 for the models with 8, 12, and 18 knots, respectively. Conversely, the estimated posterior mean of h^2_{DBH} from the model with blocks was sensibly smaller (0.08). Also, the estimate of σ^2_b from model 2 (17.351) was smaller than the estimated values from models 3 (22.317) and 4 (21.758). The estimates of σ^2_e from models 2–4 were about half the magnitude of the parameter estimate for model 1. This is due to the spatial variation not being completely accounted for by the blocking procedure in model 1. The 95% high posterior density intervals for σ^2_A, σ^2_b, σ^2_e, and h^2_{DBH} were shifted away from the zero value for all parameters. The standard errors indicate that all estimates were quite precise, although large numbers of samples were drawn to attain reasonable effective sample size (last column in Table 1).

To quantify the effect of using a single additive variance for trees with one or both parents known versus both parent unknown (bulk), we performed an analysis excluding the data of bulk trees. The posterior means of σ^2_A, σ^2_b, σ^2_e, and h^2_{DBH} were 3.914, 21.943, 10.181, and 0.277, respectively, estimates that are similar to those obtained with the entire data set.

The average accuracy of prediction of breeding values, calculated from model 3 (the one with the smallest DIC), was higher for parents (0.61) and progeny (0.54) than corresponding values (0.40 and 0.32) calculated from model 1 (Table 2). Thus, fitting a surface using B-splines resulted in a gain in accuracy of 66% for parents and 60% for offspring, a result that is due to the larger value of h^2_{DBH} estimated in the model with B-splines. The Spearman correlation between predicted breeding values from models 1 and 3 (Table 2) was 0.97 for parents and 0.94 for offspring, indicating that some reranking took place between the individuals with the least information, i.e., the progenies.

Discussion

Unaccounted for spatial variability in forest genetic trials leads to bias in estimating genetic parameters and predicting breeding values (Magnussen 1993, 1994) so that accuracy of selection decreases, thus reducing genetic gain. In the current research, we showed how to fit a two-dimensional surface using the tensor product of B-spline bases by means of a mixed model in the spirit of the P-splines of Eilers and Marx (1996, 2003). P-splines in two dimensions have also been obtained by a Bayesian approach, as shown by Lang and Brezger (2004). These authors regarded the difference matrices 3 as a first- or a second-order random walk, respectively. Our approach is different from theirs in the replacing of the singular matrix of the differences 3 by a proper variance–covariance matrix of the RKE in two dimensions. In doing so, we extend the tensor product of B-spline bases to an individual-tree mixed model to account for large-scale continuous spatial variability. Thus, the model incorporates a surface that is smoothed in the direction of both columns and rows. Gilmour at al. (1997) modeled the large-scale variation in one dimension of agricultural trials by fitting either polynomials or a cubic

© 2007 NRC Canada

Fig. 1. Spatial patterns of the residuals of tree DBH. The shading of the squares represents the magnitude of the residuals: the darker the square, the larger the residual.

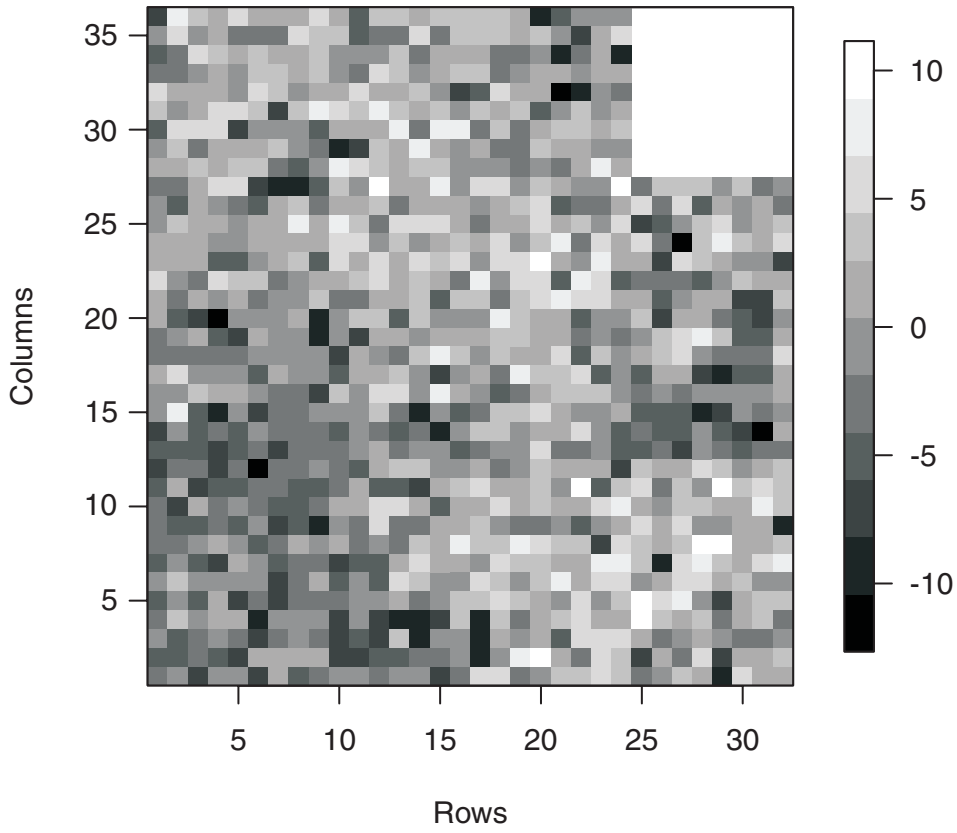
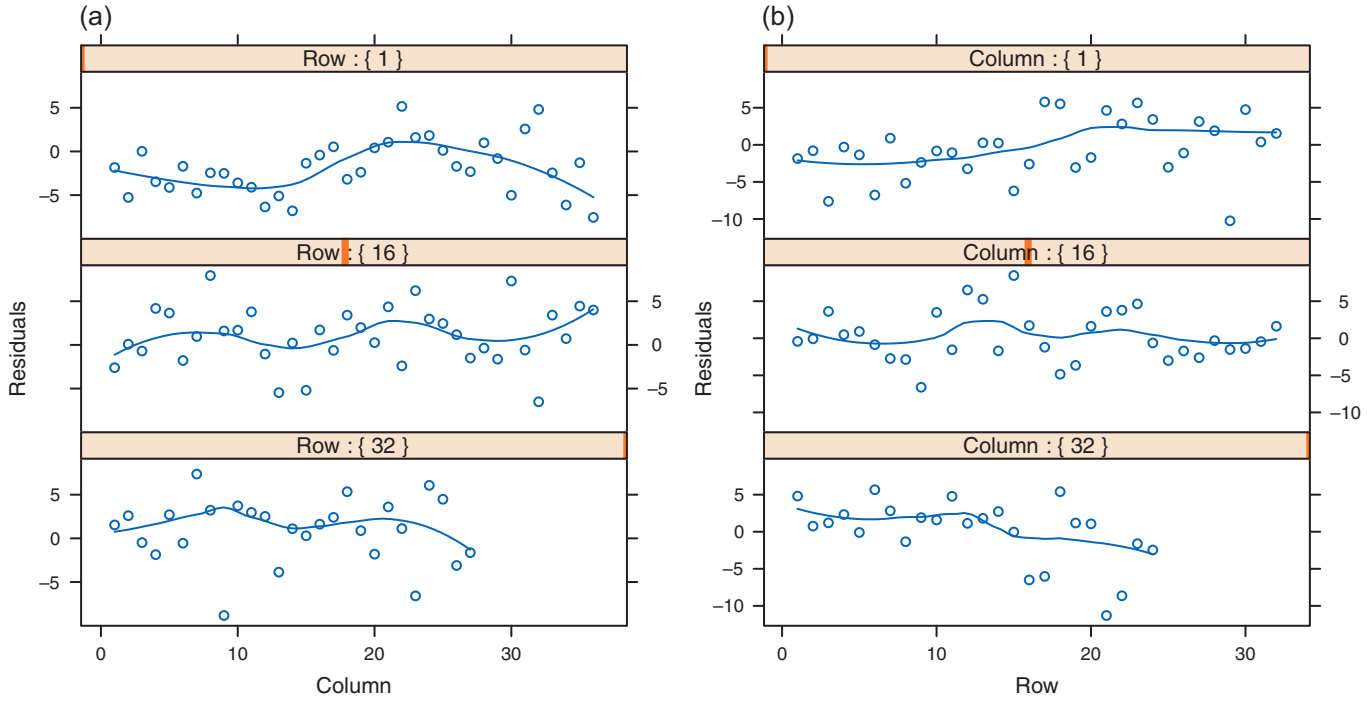


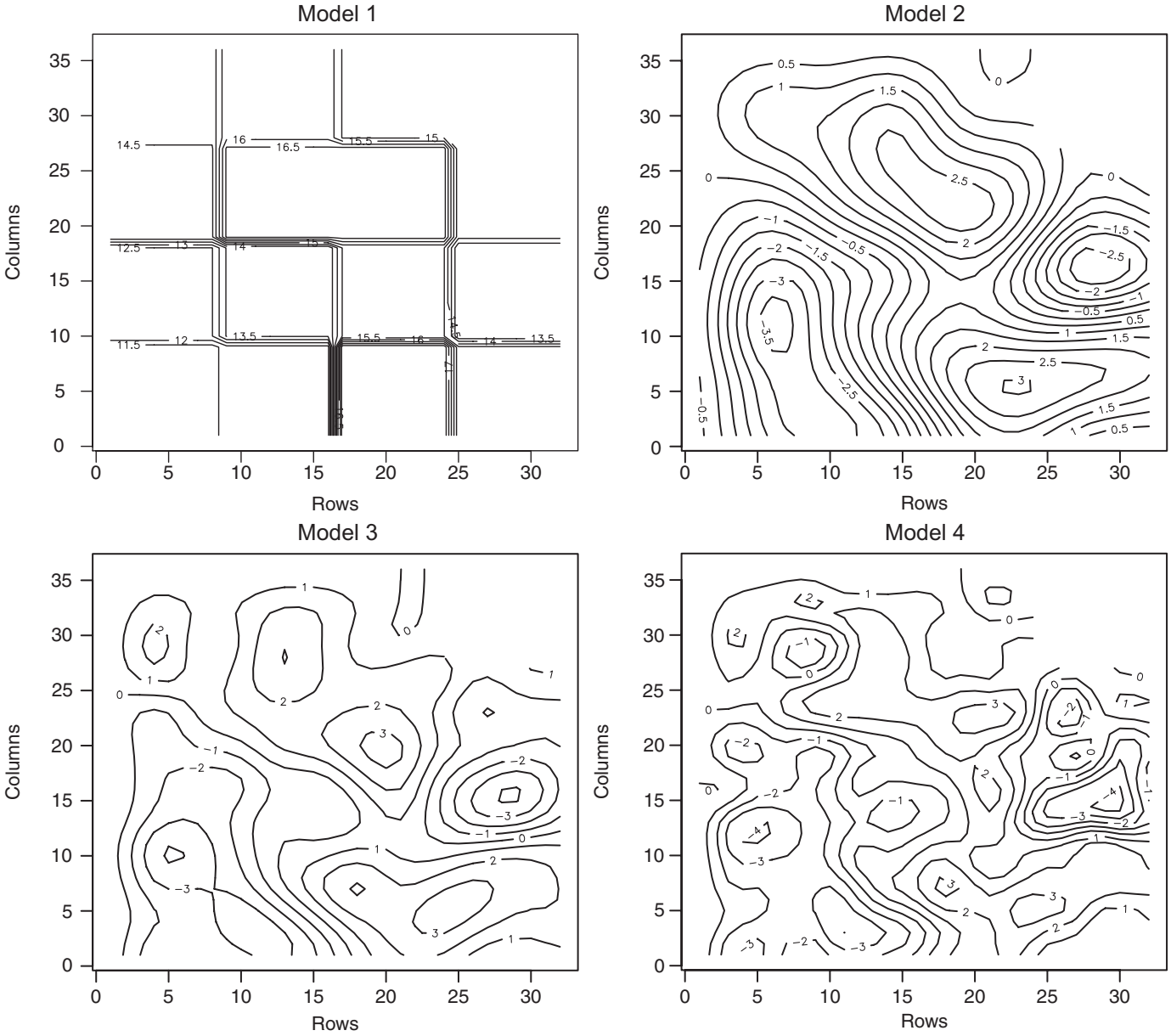
Fig. 2. Plots of the residuals after fitting provenance and additive genetic effects: (*a*) numbers of column for different rows and (*b*) number of rows for different columns.



smoothing spline. However, in forest genetic trials where trees are planted in squares or rectangles, a large portion of the global trend is usually present in the two dimensions. Moreover, it is extremely rare that large-scale continuous spatial variability is found only in the direction of the rows or of the columns, and some sort of interaction

© 2007 NRC Canada

Fig. 3. Contour plots of the estimates of block effects (model 1) and the surfaces from the fitting of the tensor product B-splines with either 8 (model 2), 12 (model 3), or 18 knots (model 4).



between rows and columns has to be considered to account for such variability (Federer 1998). Although there exist several statistical methods of smoothing to capture nonlinearity of the variation in one dimension, methods in two dimensions are less abundant. For such a purpose, Federer (1998) proposed fitting interactions between polynomials for rows and columns. However, polynomials do a poor job when fitting observations in the extremes. Moreover, small changes in the data produce a dramatic effect in the estimated values of the parameters, and this is especially so for polynomials of higher degree. Additionally, the degree of the polynomial should be selected, which in turn introduces the issue of model selection. Instead, we propose estimating a smoothed surface using P-splines. The approach is flexible, as B-spline functions are locally sensitive to the data and are numerically well conditioned. The variance σ_b^2 was used to smooth the effects of both rows and columns. In the approach of Eilers and Marx (2003) and Lang and Brezger (2004), different variances for rows and columns were used. Lang and Brezger (2004) went further and used a locally adaptive estimator of the dispersion parameters. In future research, we may consider smoothing rows and columns with different dispersion parameters, although it is not clear to us that this approach may be more advantageous than ours regarding the quality of the fit, i.e., the value of the DIC.

The P-splines methodology of Eilers and Marx (1996, 2003) consists of using cubic B-splines with equally spaced knots. In this approach, the crucial parameter is the penalty or smoothing factor λ (see eqs. 2 and 5), and the number of knots in the spline is not vital to the fit as long as there are "sufficiently" many (Eilers and Marx 1996; Cantet et al.

© 2007 NRC Canada

Table 1. Posterior statistics for the additive genetic variance (σ^2_A), the variance of the RKE (σ^2_b), the error variance (σ^2_e), and the heritability of DBH (σ^2_{DBH}).

Model	Parameter	Mean	Median	Mode	SD	95% HPD	ESS
1	σ^2_A	1.835	1.801	1.609	0.37149	1.291–2.503	24 119
	σ^2_e	23.043	20.144	14.070	8.69251	15.182–40.520	87 274
	h^2_{DBH}	0.080	0.079	0.084	0.02520	0.040–0.123	43 572
2	σ^2_A	3.596	3.480	2.642	0.98973	2.191–5.381	16 181
	σ^2_b	17.351	16.558	16.875	5.17173	10.457–26.887	169 158
	σ^2_e	11.156	11.191	10.476	1.01469	9.432–12.760	24 207
	h^2_{DBH}	0.243	0.237	0.259	0.06401	0.151–0.358	16 254
3	σ^2_A	3.754	3.643	2.933	1.00390	2.310–5.573	16 474
	σ^2_b	22.317	21.649	23.716	5.47972	14.682–32.132	109 973
	σ^2_e	10.275	10.301	9.900	1.01309	8.558–11.871	23 568
	h^2_{DBH}	0.267	0.261	0.244	0.06872	0.167– 0.389	16 519
4	σ^2_A	3.661	3.558	3.439	0.98475	2.254–5.458	16 526
	σ^2_b	21.758	21.409	18.998	4.17318	15.463–29.223	81 522
	σ^2_e	10.312	10.339	9.683	1.00670	8.595–11.920	24 305
	h^2_{DBH}	0.262	0.256	0.205	0.06706	0.164–0.383	16 588

Note: Model 1, blocks fitted as fixed effects; model 2, P-splines with 8 knots for rows and 8 knots for columns; model 3, P-splines with 12 knots for rows and 12 knots for columns; model 4, P-splines with 18 knots for rows and 18 knots for columns. HPD, high posterior density interval; ESS, effective sample size.

Table 2. Accuracy of prediction of breeding values from models 1 and 3 and Spearman correlation between predicted breeding values from models 1 and 3.

Accuracy of breeding values for parents		Accuracy of breeding values for offspring		Spearman correlation of breeding values	
Model 1	Model 3	Model 1	Model 3	Parents	Offspring
0.40	0.61	0.32	0.54	0.97	0.94

2005). In the mixed model approach to P-splines, λ is the ratio σ^2_e/σ^2_b (Cantet et al. 2005) in eq. 8. Looking at Table 1, one may infer that the magnitude of σ^2_b (the denominator of λ) was sensitive to the number of knots as compared with the other variance components. It is known that the fit of very few knots produces bias, which rapidly decreases as the number of knots increases (Ruppert 2002). Once the minimum number is reached, increasing the number of knots gives satisfactory fits (Ruppert 2002). Cantet et al. (2005) found almost equal values of the modified Akaike information criterion for models with 20, 40, 60, 80, or 120 equally spaced knots. However, restricted maximum likelihood estimators for the variance components did not converge for certain models with 120 knots. For those situations with 120 knots where convergence was attained, there were some inconsistencies in the fit for intervals where no data were recorded. It is concluded that the number of knots is not critical except for extreme quantities, and there are usually several knot numbers that yield similar fit and produce similar estimates of the variance components. In the current research, decreasing the number of knots from 18 to 8 produced a smoother surface (Fig. 3). Although the model with 12×12 knots displayed the smallest DIC, the difference in DIC between the models with 12×12 and 18×18 knots was minor. This was also true for the estimates of h^2_{DBH} obtained from both models: a difference in the third decimal place. In the mixed model approach to P-splines, the covariance structure of the RKE replaces any of the singular matrices of the differences in eq. 3. In the present research, the tridiagonal matrix proposed by Durban et al. (2001) was selected to model the covariances between the RKE for columns and for rows. The formulation is simpler than the dense correlation structures used by Cantet et al. (2005) and Hyndman et al. (2005), where there is complete dependence among all RKE. The latter covariance structures had larger DICs than the one used by Durban et al. (2001), as explained in the section entitled Models of analysis. However, similar estimates were obtained for σ^2_A (3.668, 3.753, and 3.754), for σ^2_e (10.994, 10.763, and 10.275), and for h^2_{DBH} (0.250, 0.258, and 0.267) from the models with the covariance structure used by Cantet et al. (2005), Hyndman et al. (2005), and Durban et al. (2001), respectively. On the other hand, the estimates of σ^2_b from these three models were quite different: 11.931, 1.611, and 22.317. This is in agreement with the results obtained by Cantet et al. (2005).

There are several examples of the use of B-spline functions in one dimension when analyzing breeding data. Thus, animal breeders used splines to model functional breeding values (White et al. 1999; Bohmanova et al. 2005) or the effects of management unit and time (Cantet et al. 2005). In forest genetic breeding, Cornillon et al. (2003) used B-splines to model time functional breeding values of clones in *Eucalyptus* using a fixed-effects model. Magnussen and Yanchuk (1994) fitted spline functions to observed data so as to estimate the individual heights at nonrecorded times from Douglas-fir trees. The resulting data were then used to

© 2007 NRC Canada

predict breeding values at nonrecorded ages and genetic dispersion parameters. The fit of a smoothed surface to the progeny trial in *E. globulus* subsp. *globulus* with the tensor product of B-splines instead of the a priori block design consistently increased the posterior means of σ^2_A and of h^2_{DBH} (Table 1). The results agree with those of Zas (2006) that accounted for spatial variability using kriging and are different from those of Dutkowski et al. (2002, 2006). In this latter case, inconsistent estimates of σ^2_A were obtained after adjusting an AR(1) $\times$ AR(1) covariance structure to the residuals of the model. In our data, the spatial models produced an increase in precision for the estimation of σ^2_e, which can be noticed in the much lower standard deviations and the narrower values for the 95% high posterior probability density intervals when compared with the estimate from the model with blocks (Table 1). Moreover, accuracies of breeding values from parents and offspring calculated with the spatial models were higher than corresponding values estimated from the model with block effects (Table 2) due to the increase in the estimated additive variance and the decrease of estimated error variance (Table 1). When comparing with randomized complete block designs, increases in accuracy from spatial models were reported by Costa e Silva et al. (2001) for tree height and Zas (2006) for tree diameter. Costa e Silva et al. (2001) analyzed 12 trials and found up to 71% increases in accuracy of predicted additive effects of the parents and offspring. Also, Zas (2006) reported substantial increases in the accuracy of BLUPs of family effects, going from 0.40–0.63 to 0.72–0.79 after correcting for spatially correlated variation. A smaller gain in accuracy was found by Dutkowski et al. (2002, 2006), but still in the direction of the spatial model over the model with blocks. A substantial fraction of the gain in accuracy is due to the fact that not all spatial variability is accounted for as interblock variability by using block designs (Singh et al. 2003), variation that otherwise would go to the error variance. Therefore, analysis of data displaying large-scale continuous spatial variation, such as the one induced by a petrocalcic layer at variable depth, by spatial models will most likely improve the accuracy of selection as compared with an analysis using a model with blocks.

In the current research, we modeled spatial variability that is continuous and permanent along a site using an individual-tree mixed model with a smoothed surface. In forest genetic evaluation, the spatial variation at the microsite level has been modeled with nearest neighbor techniques (Magnussen 1990; Costa e Silva et al. 2001; Dutkowski et al. 2002) or with kriging (Hamann et al. 2002; Zas 2006). Nevertheless, interplant competition may be another source for small-scale spatial variation that affects the correlation between neighbors (Magnussen 1994). The mixed model 6 does not account for genetic competition among trees, and this can bias the estimation of σ^2_A (Cappa and Cantet 2007). However, the trees used in the analysis were 6 years old so that competition was weak or absent. For those situations where trees are measured at an age where competition effects are sizeable, it would be desirable to fit simultaneously continuous spatial variation and genetic effects of competition. More important, it is worthwhile to compare the method presented in this research with other spatial techniques by computer simulation, and this is the topic of future research.

Acknowledgements

This research was supported by grants from the Agencia Nacional de Ciencia y Tecnología (FONCyT PICT 08-09502), Universidad de Buenos Aires (UBACyT G018, 2004–2007), and CONICET (PIP 5338, 2005–2006) of Argentina. The authors would like to thank INTA-Soporcel (Instituto Nacional de Tecnología Agropecuaria – Sociedad Portuguesa de Papel S.A., Portugal) for allowing us to use their data and the Associate Editor and two anonymous referees for their insightful comments on an earlier version of this manuscript.

References

Anderson, T.W. 1973. Asymptotically efficient estimation of covariance matrices with linear structure. Ann. Stat. **1**: 135–141.

Bohmanova, J., Misztal, I., and Bertrand, J.K. 2005. Studies on multiple trait and random regression models for genetic evaluation of beef cattle for growth. J. Anim. Sci. **83**: 62–67. PMID:15583043.

Cantet, R.J.C., Birchmeier, A.N., Canaza Cayo, A.W., and Fiorett, C. 2005. Semiparametric animal models via penalized splines as alternatives to models with contemporary groups. J. Anim. Sci. **83**: 2482–2494. PMID:16230644.

Cappa, E.P., and Cantet, R.J.C. 2006. Bayesian inference for normal multiple-trait individual-tree models with missing records via full conjugate Gibbs. Can. J. For. Res. **36**: 1276–1285. doi:10.1139/X06-024.

Cappa, E.P., and Cantet, R.J.C. 2007. Direct and competition additive effects in tree breeding: Bayesian estimation from an individual tree mixed model. Silvae Genet. In press. doi:10.1139/X06-024.

Christensen, R. 1987. Plane answer to complex questions. The theory of linear models. Springer-Verlag, New York.

Cornillon, P.A., Saint-Andre, L., Bouvet, J.M., Vigneron, P., Saya, A., and Gouma, R. 2003. Using B-splines for growth curve classification: applications to selection of eucalypt clones. For. Ecol. Manag. **176**: 75–85. doi:10.1016/S0378-1127(02)00276-1.

Costa e Silva, J., Dutkowski, G.W., and Gilmour, A.R. 2001. Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. Can. J. For. Res. **31**: 1887–1893. doi:10.1139/cjfr-31-11-1887.

De Boor, C. 1993. B(asic)-spline basics. Fundamental developments of computer-aided geometric modeling. *Edited by* L. Piegl. Academic Press, San Diego, Calif.

Durban, M., Currie, I., and Kempton, R. 2001. Adjusting for fertility and competition in variety trials. J. Agric. Sci. (Camb.), **136**: 129–140. doi:10.1017/S0021859601008541.

Dutkowski, G.W., Costa e Silva, J., Gilmour, A.R., and Lopez, G.A. 2002. Spatial analysis methods for forest genetic trials. Can. J. For. Res. **32**: 2201–2214. doi:10.1139/x02-111.

Dutkowski, G.W., Costa e Silva, J., Gilmour, A.R., Wellendorf, H., and Aguiar, A. 2006. Spatial analysis enhances modeling of a wide variety of traits in forest genetic trials. Can. J. For. Res. **36**: 1851–1870. doi:10.1139/X06-059.

Eilers, P.H.C., and Marx, B.D. 1996. Flexible smoothing with B-splines and penalties (with comments and rejoinder). Stat. Sci. **11**: 89–121. doi:10.1214/ss/1038425655.

Eilers, P.H.C., and Marx, B.D. 2003. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemom. Intell. Lab. Syst. **66**: 159–174.

Federer, W.T. 1998. Recovery of interblock, intergradient, and intervarietal information in incomplete block and lattice rectangle designed experiments. Biometrics, **54**: 471–481. doi:10.2307/3109756.

© 2007 NRC Canada

Fu, Y.B., Yanchuk, A.D., and Namkoong, G. 1999*a*. Incomplete block designs for genetic testing: some practical considerations. Can. J. For. Res. **29**: 1871–1878. doi:10.1139/cjfr-29-12-1871.

Fu, Y.B., Yanchuk, A.D., and Namkoong, G. 1999*b*. Spatial patterns of tree height variations in a series of Douglas-fir progeny trials: implications for genetic testing. Can. J. For. Res. **29**: 714–723. doi:10.1139/cjfr-29-6-714.

Gilmour, A.R., Cullis, B.R., and Verbyla, A.P. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. **2**: 269–293. doi:10.2307/1400446.

Green, P.J., and Silverman, B.W. 1994. Nonparametric regression and generalized linear model. Chapman and Hall, London, U.K.

Hamann, A., Koshy, M., and Namkoong, G. 2002. Improving precision of breeding values by removing spatially autocorrelated variation in forestry field experiments. Silvae Genet. **51**: 210–215.

Harville, D.A. 1997. Matrix algebra from a statistician's perspective. Springer-Verlag, New York.

Henderson, C.R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Ont.

Hobert, J.P., and Casella, G. 1996. The effects of improper priors on Gibbs sampling in hierarchical linear models. J. Am. Stat. Soc. **91**: 1461–1473.

Hyndman, R.J., King, M.L., Pitrun, I., and Billah, B. 2005. Local lineal forecasts using cubic smoothing splines. Aust. N.Z. J. Stat. **47**: 87–99. doi:10.1111/j.1467-842X.2005.00374.x.

Kass, R.E., Carlin, B.P., Gelman, A., and Neal, R.M. 1998. Markov chain Monte Carlo in practice: a roundtable discussion. Am. Stat. **52**: 93–100. doi:10.2307/2685466.

Lang, S., and Brezger, A. 2004. Bayesian P-splines. J. Comput. Graph. Stat. **13**: 183–212. doi:10.1198/1061860043010.

Libby, W.J., and Cockerham, C.C. 1980. Random non-contiguous plots in interlocking field layouts. Silvae Genet. **29**: 183–190.

Liu, J., and Burkhart, H.E. 1994. Spatial characteristics of diameter and total height in juvenile loblolly pine (*Pinus taeda* L.) plantations. For. Sci. **40**: 774–786.

Lopez, G.A., Potts, B.M., Dutkowski, G.W., and Rodriguez Traverso, J.M. 2001. Quantitative genetics of *Eucalyptus globulus*: affinities of land race and native stand localities. Silvae Genet. **50**: 244–252.

Magnussen, S. 1990. Application and comparison of spatial models in analysing tree-genetics field trials. Can. J. For. Res. **20**: 536–546.

Magnussen, S. 1993. Bias in genetic variance estimates due to spatial autocorrelation. Theor. Appl. Genet. **86**: 349–355. doi:10.1007/BF00222101.

Magnussen, S. 1994. A method to adjust simultaneously for statial microsite and competition effects. Can. J. For. Res. **24**: 985–995.

Magnussen, S., and Yanchuk, A.D. 1994. Time trends of predicted breeding values in selected crosses of coastal Douglas-fir in British Columbia: a methodological study. For. Sci. **40**: 663–685.

Marx, B.D., and Eilers, P.H. 2005. Multidimensional penalized signal regression. Technometrics, **47**: 13–22. doi:10.1198/004017004000000626.

Raftery, A.L., and Lewis, S. 1992. How many iterations in the Gibbs sampler? *In* Bayesian statistics 4. *Edited by* J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford University Press, Oxford, U.K.

Robinson, G.K. 1991. That BLUP is a good thing: the estimation of random effects. Stat. Sci. **6**: 15–32.

Ruppert, D. 2002. Selecting the number of knots for penalized splines. J. Comput. Graph. Stat. **11**: 735–757. doi:10.1198/106186002321018768.

Ruppert, D., Wand, M.P., and Carroll, R.J. 2003. Semiparametric regression. Cambridge University Press, Cambridge, U.K.

Saenz-Romero, C., Nordheim, E.V., Guries, R.P., and Crump, P.M. 2001. A case study of a provenance/progeny test using trend analysis with correlated errors and SAS PROC MIXED. Silvae Genet. **50**: 127–135.

Silverman, B. 1986. Density estimation for statistics and data analysis. Chapman and Hall, London, U.K.

Singh, M., Malhotra, R.S., Ceccarelli, S., Sarker, A., Grando, S., and Erskine, W. 2003. Spatial variability models to improve dry land field trials. Exp. Agric. **39**: 151–160. doi:10.1017/S0014479702001175.

Smith, B.J. 2003. Bayesian output analysis program (BOA). Version 1.0. User's manual. Available from www.public-health.uiowa.edu/boa/Home.html.

Sorensen, D., and Gianola, D. 2002. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer-Verlag, New York.

Soria, F., Basurco, F., Toval, G., Silió, L., Rodriguez, M.C., and Toro, M. 1998. An application of Bayesian techniques to the genetic evaluation of growth traits in *Eucalyptus globulus*. Can. J. For. Res. **28**: 1286–1294. doi:10.1139/cjfr-28-9-1286.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. 2002. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. Ser. B Methodol. **64**: 583–639. doi:10.1111/1467-9868.00353.

Thomson, A.J., and El-Kassaby, Y.A. 1988. Trend surface analysis of a Douglas-fir provenance–progeny transfer test. Can. J. For. Res. **18**: 515–520. doi:10.1139/x88-208.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., and Welham, S.J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). Appl. Stat. **48**: 269–311.

Waldmann, P., and Ericsson, T. 2006. Comparison of REML and Gibbs sampling estimates of multi-trait genetic parameters in Scots pine. Theor. Appl. Genet. **112**: 1441–1451. doi:10.1007/s00122-006-0246-x. PMID:16544126.

Wand, M.P. 2003. Smoothing and mixed models. Comput. Stat. **18**: 223–249.

White, I.M.S., Thompson, R., and Brotherstone, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. J. Dairy Sci. **82**: 632–638. PMID:10194684.

Zas, R. 2006. Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree Genet. Genomes, **2**: 177–185. doi:10.1007/s11295-006-0042-4.

© 2007 NRC Canada