Analysis of Longitudinal Data from Progeny Tests: Some Multivariate Approaches

Luis A. Apiolaza and Dorian J. Garrick

ABSTRACT. Longitudinal data arise when trees are repeatedly assessed over time. The degree of genetic control of tree performance typically changes over time, creating relationships between breeding values at different ages. Longitudinal data allow modeling the changes of heritability and genetic correlation with age. This article presents a tree model (i.e., a model that explicitly includes a term for additive genetic effects of individual trees) for the analysis of longitudinal data from a multivariate perspective. The additive genetic covariance matrix for several ages can be expressed in terms of a correlation matrix pre- and post-multiplied by a diagonal matrix of standard deviations. Several models to represent this correlation matrix (unstructured, banded correlations, autoregressive, full-fit and reduced-fit random regression, repeatability, and uncorrelated) are presented, and the relationships among them explained. Kirkpatrick's alternative approach for the analysis of longitudinal data using covariance functions is described, and its similarities with the other models discussed in this article are detailed. The use of Akaike's information criterion for model selection considering likelihood and number of parameters is detailed. All models are illustrated through the analysis of weighed basic wood density (in kg/m^3) at four ages (5, 10, 15, and 20 yr) from radiata pine increment cores. FOR. SCI. 47(2):129–140.

Key Words: Multivariate analysis, tree model, covariance structures, covariance functions, BLUP.

TREE BREEDING HAS A MULTIVARIATE NATURE. In most breeding programs, the selection criteria involve two or more characteristics. Apart from the obvious use when dealing with different traits (e.g., growth and wood properties), a multivariate approach can be utilized with different expressions of the same trait. Hence, problems of a seemingly univariate structure can be fully exploited in a multivariate framework. For example, growth rate assessed in two different environments can be modeled as if controlled by different genes, and treated as a multivariate analysis (Falconer 1952). Here the genetic correlation between the traits is a measure of genotype by environment interaction. Another application, which we study here, is in the analysis of longitudinal data that arise when trees are repeatedly assessed at several points in time (e.g., basic wood density at ages 5, 10, and 15). Thus, expressions of the trait at different times are considered different variables.

We make a distinction between longitudinal data and repeated measures because the latter term not only includes different times (longitudinal data) but also multiple assessments of morphological traits (e.g., lengths of right and left wings of a bird) or measures under different conditions (Cnaan et al. 1997). Longitudinal data can be considered a particular form of multivariate data—because the "same trait" is measured at each time, there is an underlying continuum (time) and the sequential nature of measurement creates patterns of variation (Hand and Crowder 1996).

Dr. Luis A. Apiolaza is Quantitative Geneticist, Cooperative Research Centre for Sustainable Production Forestry, School of Plant Science, University of Tasmania, GPO Box 252-55, Hobart, Tasmania 7001, Australia—Phone: (+61) 3 6226 2213; Fax: (+61) 3 6226 2698; E-mail: luis.apiolaza@utas.edu.au. When writing this paper he was a Ph.D. student at the Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand and New Zealand Forest Research Institute, Rotorua, New Zealand. Prof. Dorian J. Garrick holds the A.L. Rae Chair in Animal Breeding and Genetics, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand—Phone: (+64) 6 350 5103; E-mail: d.garrick@massey.ac.nz.

Acknowledgments: L. Apiolaza was funded by NZODA and NZFRI scholarships. The dataset used in the example was compiled by Paul Jefferson, based on results from a densitometry analysis. The NZ Radiata Pine Breeding Cooperative kindly provided the densitometry results. Comments by Rowland Burdon (New Zealand Forest Research Institute), Mark Dieters (Queensland Forest Research Institute), Tore Ericsson (SkogForst), Nicolás López-Villalobos (Massey University), and the anonymous reviewers improved the original manuscript.

Manuscript received August 8, 1999. Accepted May 16, 2000.

Longitudinal data allow modeling the changes of heritability and genetic correlations with age. Therefore, data from multiple assessments may be integrated in the prediction of breeding values and this allows the evaluation time for early selection to be optimized (Burdon 1989). Longitudinal data are a frequent feature of tree breeding programs; however, their analysis has often been reduced to a univariate approach. There are examples of multivariate modeling of longitudinal data in forest mensuration (e.g., Gregoire et al. 1995). Multivariate applications in tree breeding are scarce and have typically considered only a full unstructured approach (e.g., Wei and Borralho 1998). The only exception we are aware of is Magnussen and Kremer (1993), fitting growth models to individual trees and Apiolaza et al. (2000), comparing different parameterizations of the additive genetic covariance matrices. Although the use of best unbiased linear prediction and tree models (Henderson 1984) is increasingly popular (e.g., Borralho 1995), there is no unified presentation of its theoretical background and the link between univariate and multivariate analyses in a tree breeding context. Furthermore, simple models like covariance functions, well known in evolutionary genetics and animal breeding, have received little attention in tree breeding, and their relationship with multivariate analysis has not yet been discussed.

This article provides a unified presentation of multivariate analysis with longitudinal data from progeny trials (i.e., with a genetic structure) using a tree model. A univariate tree model is detailed and then extended to multivariate form. We explain the concept of covariance structures and show the relationships among these structures and the corresponding predicted breeding values. Several statistical models to deal with covariance structures are specified, the relationship between full multivariate analysis and random regression models is demonstrated, and model selection techniques are presented. An alternative approach, covariance functions, is also discussed. An example is developed comparing the different models.

Univariate Analysis

In a typical univariate analysis the scalar phenotypic observation y_i on individual i is expressed in the so-called tree model (see Borralho 1995) as a function of fixed effects, additive genetic value of the tree (a_i) and a residual effect (e_i):

$$y_i = \mathbf{x}_i' \mathbf{b} + a_i + e_i \tag{1}$$

where $\mathbf{y}$ is a vector of observations on one trait, $\mathbf{b} = [b_1\ b_2 \ldots b_f]'$ is the vector of fixed effects (e.g., overall mean, site, etc.) and $\mathbf{x}_i' = [1 \ldots]$ is a row vector containing 1's and 0's linking observations to the fixed effects. This notation is for the observation of a single individual. Considering all N trees under analysis, and extending the matrix notation, Equation (1) becomes:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{a} + \mathbf{e} \tag{2}$$

where $\mathbf{b}$ is the vector of fixed effects (as defined before), $\mathbf{a} = [a_1\ a_2 \ldots a_N]'$ is the vector of random additive genetic values, and $\mathbf{e} = [e_1\ e_2 \ldots e_N]'$ is the vector of random residuals. The incidence matrices $\mathbf{X}$ (obtained by stacking $\mathbf{x}_i'$ for all trees) and $\mathbf{Z}$ links observations to $\mathbf{b}$ and $\mathbf{a}$, respectively. The vector of expected values and the dispersion matrices are defined by:

$$\begin{aligned} &\mathrm{E}[\mathbf{y}] = \mathbf{X}\,\mathbf{b} \\ &\mathrm{Var}[\mathbf{a}] = \mathbf{G} = \mathbf{A}_N\sigma_a^2,\ \mathrm{Var}[\mathbf{e}] = \mathbf{R} = \mathbf{I}\sigma_e^2 \\ &\text{and } \mathrm{Var}[\mathbf{y}] = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \end{aligned} \tag{3}$$

where $\mathbf{A}_N$ is the numerator relationship matrix, which describes the additive genetic relationship among individuals (see Mrode 1996, Chapter 2, for a detailed explanation). In addition, $\mathbf{I}$ is an identity matrix, σ_a^2 is the additive genetic variance, and σ_e^2 is the error variance. Random effects $\mathbf{a}$ and $\mathbf{e}$ are assumed to be uncorrelated.

The analysis of progeny tests normally involves two steps: first the estimation of variance components and second the prediction of breeding values for the individuals, using the variance components estimated in the first step. Restricted maximum likelihood (REML, Patterson and Thompson 1971) is being increasingly used for variance components estimation in tree breeding (e.g., Huber et al. 1994, Dieters et al. 1995), although there are now a few applications with a Bayesian approach using Monte Carlo Markov Chains (e.g., Soria et al. 1997).

Assuming that $\mathbf{y}$, $\mathbf{a}$, and $\mathbf{e}$ follow a multivariate normal distribution, and provided $\mathbf{G}$ and $\mathbf{R}$ are positive definite, best linear unbiased predictions (BLUP), (Henderson 1984) of the breeding values of the individuals are calculated using Henderson's mixed model equations (Henderson 1984):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{4}$$

where $\mathbf{G}$ and $\mathbf{R}$ are functions of σ_a^2 and σ_e^2 respectively [see Equation (3)]. In practice, estimates $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$ are used in place of unknown parameters, so the predicted breeding values are in fact approximations of BLUP.

To obtain REML estimates of variance components the log-likelihood (Log L) function is maximized with respect to σ_a^2 and σ_e^2, subject to the constraints that these parameters are within the parameter space (i.e., nonnegative and less or equal to the total phenotypic variance):

$$\begin{aligned} &Log\ L = \\ &-1/2\ [con + \log|\mathbf{G}| + \log|\mathbf{R}| + \log|\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y}] \end{aligned} \tag{5}$$

where *con* is a constant, $\mathbf{G}$ and $\mathbf{R}$ are as from Equation (3), $\mathbf{C}$ is the coefficient matrix of Equation (4), $\mathbf{P}$ is the projection matrix $\mathbf{V}^{-1} - \mathbf{V}^{-1}\,\mathbf{X}\,(\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X})^{-}\mathbf{X}'\,\mathbf{V}^{-1}$, and $(\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X})^{-}$ represents a generalized inverse of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$. The matrix $\mathbf{P}$ absorbs the fixed effects and accounts for information about $\mathbf{V}$.

Multivariate Analysis

The steps involved in a multivariate analysis are similar to the univariate case. Consider now a vector $\mathbf{y}_i = [y_{i1}\ y_{i2} \ldots y_{im}]'$ representing m observations (either different traits or

repeated measurements) on individual *i*. This vector of phenotypic observations can be expressed in terms of genetic and environmental components using:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{a}_i + \mathbf{e}_i \quad (6)$$

where $\mathbf{b} = [\mathbf{b}'_{\text{trait1}}\ \mathbf{b}'_{\text{trait2}} \ldots \mathbf{b}'_{\text{trait}m}]'$ is the vector of fixed effects (which can be different for each trait), $\mathbf{a}_i = [a_{i1}\ a_{i2} \ldots a_{im}]'$ is the vector of random additive genetic effects and $\mathbf{e}_i = [e_{i1}\ e_{i2} \ldots e_{im}]'$ is the vector of random residuals. The incidence matrices have the same function as in the univariate case, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ have one row for each observation in $\mathbf{y}_i$. Note the use of matrix notation for additive genetic effects and residuals already at the individual level, and the similarity to Equation (2) (but for the subscript *i*).

The expected value and dispersion for a noninbred individual are defined by:

$$E[\mathbf{y}_i] = \mathbf{X}_i\,\mathbf{b}$$

$$\begin{aligned} &\text{Var}[\mathbf{a}_i] = \mathbf{G}_0, \text{Var}[\mathbf{e}_i] = \mathbf{R}_0 \\ &\text{and} \\ &\text{Var}[\mathbf{y}_i] = \mathbf{Z}_i\mathbf{G}_0\mathbf{Z}_i' + \mathbf{R}_0 \end{aligned} \quad (7)$$

In the multivariate approach, $\mathbf{G}_0$ and $\mathbf{R}_0$ represent the $m \times m$ additive genetic and residuals covariance matrices between the observations, respectively. Their typical elements for traits (or measurements) *j* and *k* are $\sigma_{a_{jk}}$ and $\sigma_{e_{jk}}$. Again, random effects $\mathbf{a}_i$ and $\mathbf{e}_i$ are assumed uncorrelated. This model can be easily expanded to include more random effects such as block and plot effects (see, for example, Apiolaza et al. 2000).

This multiple-trait model for one individual is extended to the *N* individuals in the progeny test using Equation (2), but now $\mathbf{y} = [\mathbf{y}_1' \mathbf{y}_2' \ldots \mathbf{y}_N']'$, $\mathbf{a} = [\mathbf{a}_1' \mathbf{a}_2' \ldots \mathbf{a}_N']'$ and $\mathbf{e} = [\mathbf{e}_1' \mathbf{e}_2' \ldots \mathbf{e}_N']'$. In addition, $\mathbf{X} = [\mathbf{X}_1' \mathbf{X}_2' \ldots \mathbf{X}_N']'$ and $\mathbf{Z} = \sum_{\oplus} \mathbf{Z}_i$, where $\sum_{\oplus}$ represents direct sum operation. Consequently, $\mathbf{G} = \mathbf{A}_N \otimes \mathbf{G}_0$ and $\mathbf{R} = \sum_{\oplus} \mathbf{R}_i$, where $\otimes$ denotes direct product [see Appendix 1 and Searle 1982 (Chapter 10) for a detailed description of $\sum_{\oplus}$ and $\otimes$ operations] and $\mathbf{R}_i$ is the residual covariance matrix for each individual. Hence, the expected value and dispersion matrices are:

$$\text{E}[\mathbf{y}] = \mathbf{X}\,\mathbf{b}$$

$$\begin{aligned} &\text{Var}[\mathbf{a}] = \mathbf{G} = \mathbf{A}_N \otimes \mathbf{G}_0, \text{Var}[\mathbf{e}] = \mathbf{R} = \textstyle\sum_{\oplus} \mathbf{R}_i \\ &\text{and} \\ &\text{Var}[\mathbf{y}] = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \end{aligned} \quad (8)$$

Once the model is defined, the analysis of the multivariate expression of Equations (4) and (5) is developed in ways similar to the univariate estimation of variance parameters and to predict breeding values.

Analysis of Longitudinal Data: Covariance Structures

The use of multivariate models with unstructured covariance matrices (i.e., not assuming any patterns) for the analysis of *m* repeated measurements is an appropriate, but not necessarily the best, option. Each of these covariance matrices involves the estimation of $m(m + 1)/2$ covariance components. In comparison to a univariate analysis, the amount of data on each subject increases by *m*, but the number of covariance parameters to estimate increases by $m(m + 1)/2$. Therefore the information available to estimate each parameter is in some sense reduced, as may be the "quality" of the estimates. Modeling the covariance structures reduces the number of parameters to estimate and can provide explanation for patterns of observed correlation among the longitudinal data.

Covariance matrices (**M**) can generally be expressed as a symmetric correlation matrix (**C**) with typical element r_{jk} pre- and post-multiplied by a diagonal matrix (**S**) containing the square root of the variance components for each trait (measure). Hence:

$$\mathbf{M} = \mathbf{S}\,\mathbf{C}\,\mathbf{S} \quad (9)$$

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix} \quad (10)$$

This notation simplifies the explanation of the structures used for modeling the covariance matrices. We typically allow heterogeneous variances in time, so **S** is a diagonal matrix with all diagonal elements different. In case of stable processes, or stabilized through transformation to a homogeneous variance, $\mathbf{S} = \mathbf{I}\,\sigma$, a diagonal with identical elements. Below we provide a list of some common, but not exhaustive, structures for **C**, where scalars denoted with different letters represent different correlations. Each structure is followed by the relationship between successive predicted breeding values. While structures can be applied to **G** and **R**, in this article we emphasize modeling the additive genetic covariance matrix, while keeping the residuals matrix unstructured. The only exceptions are the repeatability and uncorrelated models. All examples consider four measurements.

Unstructured (US)

The unstructured model can be expressed as $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{US}\,\mathbf{S}$, where $\mathbf{C}_{US}$ have no restrictions except for being positive definite and with elements between –1 and 1. This is the choice when working with different variables. Its main problem with longitudinal data is the risk of overparameterization, with poorly estimated parameters and maybe unnecessary computational requirements.

$$\mathbf{C}_{US} = \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix} \quad (11)$$

The breeding value of individual *i* observed at time *j* (a_{ij}) is a function of genes involved in expression at time

$j - k$ ($a_{i\,j-k}$) plus the effect of genes acting in the new measurement (α_j), which are considered independent of the past measurement:

$$a_{ij} = \rho_{jk}\, a_{i\,j-k} + \alpha_j \tag{12}$$

where ρ_{jk} is the additive genetic correlation between measures j and k, and $j - k \geq 0$.

Banded Correlations (BC)

The banded correlations model accommodates the existence of identical correlations for measurements with the same time between expressions (lag). Thus $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{BC}\,\mathbf{S}$, with $\{a, d, f\}\rightarrow g$, $\{b, e\}\rightarrow h$, and $\{c\}\rightarrow i$ respectively from Equation (11) ($\mathbf{C}_{US}$). If the lag between all measures is the same, the correlation matrix presents bands with the same value [see Equation (13)].

$$\mathbf{C}_{BC} = \begin{bmatrix} 1 & g & h & i \\ g & 1 & g & h \\ h & g & 1 & g \\ i & h & g & 1 \end{bmatrix} \tag{13}$$

The relationship between successive breeding values is similar to Equation (12), but ρ is the same for all observations separated by a lag k:

$$a_{ij} = \rho_k\, a_{i\,j-k} + \alpha_j \tag{14}$$

This assumption may not be applicable across different growth stages, where development in 1 yr of, say, early growth can be very different from that of 1 yr in mature growth (due to ontogenetic effects).

Autoregressive (AR)

Rather than using a different correlation for each lag, the autoregressive model postulates a mechanism where the correlation between measurements j and k is $r^{|j-k|}$. In this model $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{\mathrm{AR}}\,\mathbf{S}$, further reducing to 1 the number of covariances to estimate.

$$\mathbf{C}_{AR} = \begin{bmatrix} 1 & a^{|t_2-t_1|} & a^{|t_3-t_1|} & a^{|t_4-t_1|} \\ a^{|t_2-t_1|} & 1 & a^{|t_3-t_2|} & a^{|t_4-t_2|} \\ a^{|t_3-t_1|} & a^{|t_3-t_2|} & 1 & a^{|t_4-t_3|} \\ a^{|t_4-t_1|} & a^{|t_4-t_2|} & a^{|t_4-t_3|} & 1 \end{bmatrix} \tag{15}$$

Again, the breeding value of individual i observed on time j (a_{ij}) is a function of genes acting at time j–1 ($a_{i\,j-1}$) plus genes acting on the new measurement (α_j):

$$a_{ij} = \rho a_{ij-1} + \alpha_j \tag{16}$$

if the correlation (ρ) is a function of a unique value and the lag between the measurements, the relationship between successive breeding values for individual i is:

$$\begin{aligned} a_{ij-1} &= \rho a_{ij-2} + \alpha_{j-1} \\ a_{ij-2} &= \rho a_{ij-3} + \alpha_{j-2} \\ &\vdots \\ a_{ij-k+1} &= \rho\, a_{ij-k} + \alpha_{j-k+1} \end{aligned} \tag{17}$$

and substituting every breeding value of Equation (17) in the preceding one we obtain:

$$a_{ij} = \rho^{|j-k|} a_{ij-k} + \alpha_j' \tag{18}$$

where α_j' represents genes acting on measurement j plus a series of lag effects from previous innovation terms.

The autoregression coefficient can have a power formulation as $\rho = e^{-k\ \mathrm{lag}}$ (Diggle 1988) allowing for analysis with unequally spaced observations. This model is appropriate for smooth changes of genetic correlations with time, and the presence of smaller correlations at the initial stages of a trial can sometimes be modeled changing the units of the time scale (e.g., to natural logarithm or square root). A generalization of the autoregressive model is *ante-dependence*, where the breeding value is a function of n previous breeding values (Gabriel 1962).

Repeatability (RE)

This model considers longitudinal data as expressions of the same trait (under control of the same genes); that is, a genetic correlation of 1 is assumed, with homogeneous heritability on time, and equal environmental correlation between all pairs of records. Thus, $\mathbf{G}_0 = \sigma_a^2\mathbf{J}$ and $\mathbf{R}_0 = \sigma_e^2(\mathbf{I} + \rho\mathbf{J})$, where $\mathbf{J}$ is a square matrix with all elements equal to 1 and $\rho/(1 + \rho)$ is the correlation between residuals. Therefore $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{RE}\,\mathbf{S}$, with $\mathbf{S} = \mathbf{I}\,\sigma_a$ and $\mathbf{C}_{RE} = \mathbf{J}$.

$$\mathbf{C}_{RE} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{19}$$

As all rows are identical, $\mathbf{G}_0$ is singular, impeding the use of mixed model equations [Equation (4)] in its normal form. A solution for this problem is the regularly used alternative "univariate" representation of the model:

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{a} + \mathbf{W}\,\mathbf{h} + \mathbf{e} \tag{20}$$

that is, an extension of Equation (2) (univariate analysis) where $\mathbf{h} = [h_1\ h_2\ \ldots\ h_N]'$ is a vector of "permanent environmental effects," which takes into account the residual covariance between measurements, and $\mathbf{W}$ an incidence matrix. Additive genetic variance $(\mathbf{G} = \mathbf{A}_N\sigma_a^2)$ and residuals variance $(\mathbf{R} = \mathbf{I}\sigma_e^2)$ are like in the univariate case, while phenotypic variance now includes permanent environment variance:

$$\mathrm{E}[\mathbf{y}] = \mathbf{X}\,\mathbf{b}$$

$$\mathrm{Var}[\mathbf{h}] = \mathbf{H} = \mathbf{I}\sigma_h^2 \text{ and } \mathrm{Var}[\mathbf{y}] = \mathbf{ZGZ}' + \mathbf{WHW}' + \mathbf{R} \tag{21}$$

A common problem is scale difference between measures. However, this difference may be avoided using a transformation for stabilizing variance (e.g., logarithmic, Box-Cox, etc.). Nevertheless, with tree breeding experiments spanning several years (even decades), the equal correlation assumptions are sometimes naïve. In spite of this, the RE model could be useful for some short-term experiments.

Uncorrelated (UC)

The uncorrelated model assumes that there is no genetic and no residual association between successive observations. Thus $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{UC}\,\mathbf{S}$, where $\mathbf{C}_{UC} = \mathbf{I}$, an identity matrix. This is equivalent to univariate analysis by age, allowing the calculation of heritabilities but not of correlations between measures.

$$\mathbf{C}_{UC} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{22}$$

This model may be adequate when all trees are measured at all times, but it is not appropriate in the presence of selection (thinnings, mortality, etc.) whereby remaining individuals are a selected sample based on performance at earlier ages.

Random Regressions (RRf and RRr)

The phenotypic trajectory of a trait (dependent on time) can be expressed through a mathematical function tractable in a mixed linear model framework, for example, using polynomial regression, growth models, or cubic splines. A general representation for the measurements of individual i might be:

$$\mathbf{y}_i = \mathrm{f}_{\mathbf{b}}(\mathbf{t}) + \mathrm{f}_{\mathbf{a}_i}(\mathbf{t}) + \mathrm{f}_{\mathbf{e}_i}(\mathbf{t}) + \varepsilon_i \tag{23}$$

where $\mathrm{f}_{\mathbf{b}}(\mathbf{t})$, $\mathrm{f}_{\mathbf{a}_i}(\mathbf{t})$ and $\mathrm{f}_{\mathbf{e}_i}(\mathbf{t})$ represent possibly different functions modeling fixed effects, additive genetic effects and residuals respectively; and ε_i is an error term. Functions can be applied to all components of the phenotype (e.g., fixed effects, tree, and residuals) or to specific elements (e.g., tree only). Again, the emphasis is on modeling the additive genetic covariance matrix ($\mathbf{G}$), with random regressions used for $\mathbf{a}_i$ while other terms are considered unstructured and the subindex for $f(\mathbf{t})$ is dropped. If $\mathbf{a}_i = f(\mathbf{t})$ with $\mathbf{t}$ a vector of times, rather than estimating one breeding value for each assessment, the coefficients of a function that models the trajectory are estimated. Consider, for purposes of illustration, an orthogonal polynomial function to model the breeding value of individual i on time j (a_{ij}):

$$a_{ij} = f(t_j) = \lambda_{0i} z_{0j} + \lambda_{1i} z_{1j} + \lambda_{2i} z_{2j} + \ldots + \lambda_{ni} z_{nj} \tag{24}$$

where λ_{ki} are the random regression coefficients, z_{kj} is the kth orthogonal polynomial evaluated at age j, and $n \leq m-1$. Thus, all breeding values of individual i can be represented as:

$$\mathbf{a}_i = f(\mathbf{t}) = \mathbf{Q}_i\,\lambda_i \tag{25}$$

where $\lambda_i = [\lambda_{0i}\ \lambda_{1i} \ldots \lambda_{ni}]'$ and $\mathbf{Q}_i$ is an incidence matrix of form:

$$\mathrm{Q}_i = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{n1} \\ z_{02} & z_{12} & \cdots & z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0m} & z_{1m} & \cdots & z_{nm} \end{bmatrix} \tag{26}$$

Therefore, Equation (6) can be represented as:

$$\mathbf{y}_i = \mathbf{X}_i\,\mathbf{b} + \mathbf{Q}_i\,\lambda_i + \mathbf{e}_i \tag{27}$$

with

$$\mathrm{Var}[\mathbf{Q}_i\,\lambda_i] = \mathbf{Q}_i\,\Lambda_0\,\mathbf{Q}_i' \tag{28}$$

where Λ_0 is the covariance matrix of the random coefficients (λ_i). Because different regression coefficients are calculated for every individual (and these coefficients are considered as random effects), this model is called the "random regressions model." When the polynomial is of maximum degree ($m - 1$), there is a full fit (*RRf*), that is, the function $f(\mathbf{t})$ goes through all the points/measurements. In this case, the estimates using $f(\mathbf{t})$ are equivalent to those using a full multivariate approach (see below). A polynomial of order lower than $m - 1$ generates a reduced fit (*RRr*) and, in fact, is smoothing the covariance matrix.

Including polynomials evaluated at additional ages in $\mathbf{Q}_i$, within the age range used to generate the function, interpolates the appropriate covariances. Extrapolating covariances outside the range used for constructing the function is possible; however, there are no provisions in the method to ensure reliable prediction of the covariances.

Further details of these models can be found in Laird and Ware (1982, RR); Quaas et al. (1984 p. 34, RE); Jennrich and Schluchter (1986, US, BC, AR, RR, and UC); Louis (1988, RR); Diggle et al. (1994, RR); Everitt (1995, RR); Hand and Crowder (1996, US, AR, and RR); and Cnaan et al. (1997, RR). Diggle et al. (1994, Chapter 5) and Hand and Crowder (1996, Chapter 6) provide an extensive treatment of the topic.

Relationship Between Unstructured and Random Regression Models

Two linear models, m_1 and m_2, are considered equivalent when their expected values and variances are identical (Henderson 1984, p. 6):

$$E[m_1] = E[m_2]$$

$$\mathrm{Var}[m_1] = \mathrm{Var}[m_2] \tag{29}$$

The equivalency between the US and full fit RR models and the relationship between US and reduced fit RR models will be illustrated with an example. Suppose a progeny test was assessed four times (see Figure 1). We present a set of observations for a generic individual according to the model in 6 and 7. We have no particular interest in the fixed effects, which will be represented as $\mathbf{X}_i\,\mathbf{b}$.

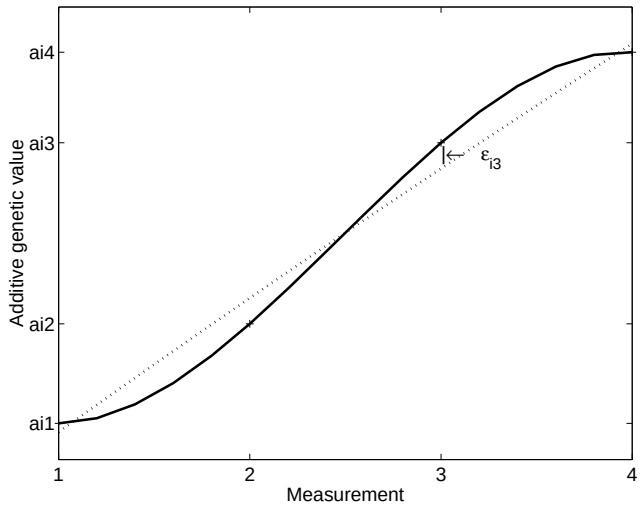

Figure 1. Fitting four measurements using the Full-fit Random Regression model (RRf) from Equation (32) (—) and the Reduced-fit Random Regression model (RRr) from Equation (34) (····). The error in estimating the additive genetic value for measurement 3, due to fitting a reduced model, is represented by ε_{i3}.

Using a US multivariate approach [i.e., model Equations (6) and (11), where $\mathbf{a}_i$ is the vector of additive values at different measurements times], we get:

$$\mathbf{y}_i = \mathbf{X}_i\,\mathbf{b} + \mathbf{Z}_i\,\mathbf{a}_i + \mathbf{e}_i, \text{ or}$$

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \mathbf{X}_i\mathbf{b} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \\ e_{i4} \end{bmatrix} \quad (30)$$

Using a full fit polynomial regression (*RRf*) [i.e., model Equation (27), where λ_i represents the regression coefficients, to model the additive genetic part] we have:

$$\mathbf{y}_i = \mathbf{X}_i\,\mathbf{b} + \mathbf{Q}_i\,\lambda_i + \mathbf{e}_i, \text{ or}$$

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \mathbf{X}_i\mathbf{b} + \begin{bmatrix} z_{01} & z_{11} & z_{21} & z_{31} \\ z_{02} & z_{12} & z_{22} & z_{32} \\ z_{03} & z_{13} & z_{23} & z_{33} \\ z_{04} & z_{14} & z_{24} & z_{34} \end{bmatrix} \begin{bmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \\ \lambda_{3i} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \\ e_{i4} \end{bmatrix} \quad (31)$$

Because a polynomial of degree $n-1$ will pass through all n observations (Neter and Wasserman 1974, p. 276), the product $\mathbf{Q}_i\,\lambda_i$ on Equation (31) is:

$$\begin{aligned} \lambda_{0i}z_{01} + \lambda_{1i}z_{11} + \lambda_{2i}z_{21} + \lambda_{3i}z_{31} &= a_{i1} \\ \lambda_{0i}z_{02} + \lambda_{1i}z_{12} + \lambda_{2i}z_{22} + \lambda_{3i}z_{32} &= a_{i2} \\ \lambda_{0i}z_{03} + \lambda_{1i}z_{13} + \lambda_{2i}z_{23} + \lambda_{3i}z_{33} &= a_{i3} \\ \lambda_{0i}z_{04} + \lambda_{1i}z_{14} + \lambda_{2i}z_{24} + \lambda_{3i}z_{34} &= a_{i4} \end{aligned} \quad (32)$$

that is also the result of the product $\mathbf{Z}_i\,\mathbf{a}_i$ in Equation (30). If $\mathbf{Z}_i\,\mathbf{a}_i$, and $\mathbf{Q}_i\,\lambda_i$ are identical, so are their variances. The expected values for both Equations (30) and (31) are $\mathbf{X}_i\,\mathbf{b}$. Thus,

$$E[US] = E[RRf] = \mathbf{X}_i\,\mathbf{b}$$

$$\text{Var}[US] = \text{Var}[RRf] = \mathbf{Z}_i\,\mathbf{G}_0\,\mathbf{Z}_i' + \mathbf{R}_0 = \mathbf{Q}_i\,\Lambda_0\,\mathbf{Q}_i' + \mathbf{R}_0 \quad (33)$$

and the models are equivalent. Moreover, random regression coefficients can be estimated from the US model as $\lambda_i = \mathbf{Q}_i^{-1}\mathbf{Z}_i\mathbf{a}_i = \mathbf{Q}_i^{-1}\mathbf{a}_i$.

Using a reduced fit, for example a quadratic polynomial (Figure 1), we have:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \mathbf{X}_i\mathbf{b} + \begin{bmatrix} z_{01} & z_{11} & z_{21} \\ z_{02} & z_{12} & z_{22} \\ z_{03} & z_{13} & z_{23} \\ z_{04} & z_{14} & z_{24} \end{bmatrix} \begin{bmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{bmatrix} + \begin{bmatrix} e^*_{i1} \\ e^*_{i2} \\ e^*_{i3} \\ e^*_{i4} \end{bmatrix} \quad (34)$$

Because the reduced fit polynomial will not in general fit the four observations perfectly we have that:

$$\begin{aligned} \lambda_{0i}z_{01} + \lambda_{1i}z_{11} + \lambda_{2i}z_{21} + \varepsilon_{i1} &= a_{i1} \\ \lambda_{0i}z_{02} + \lambda_{1i}z_{12} + \lambda_{2i}z_{22} + \varepsilon_{i2} &= a_{i2} \\ \lambda_{0i}z_{03} + \lambda_{1i}z_{13} + \lambda_{2i}z_{23} + \varepsilon_{i3} &= a_{i3} \\ \lambda_{0i}z_{04} + \lambda_{1i}z_{14} + \lambda_{2i}z_{24} + \varepsilon_{i4} &= a_{i4} \end{aligned} \quad (35)$$

where $\varepsilon_i = [\varepsilon_{i1}\ \varepsilon_{i2}\ \varepsilon_{i3}\ \varepsilon_{i4}]'$ is the vector containing the errors due to fitting a reduced regression model for the additive genetic effects. Thus, $\mathbf{e}^*_i = \mathbf{e}_i + \varepsilon_i$. In other words, the error of the full fit model ($\mathbf{e}_i$) plus the error due to the regression model (ε_i) compound a new error $\mathbf{e}^*_i$. Figure 1 depicts the difference between fitting a full-fit and a reduced-fit random regression model, and the graphical meaning of ε_i.

The expected value of the model is still the same ($\mathbf{X}_i\,\mathbf{b}$), but the dispersion matrices are now:

$$\text{Var}[\lambda_i] = \Lambda_0,\ \text{Var}[\mathbf{e}^*_i] = \mathbf{R}^*_0 \text{ and } \text{Var}[\mathbf{y}_i] = \mathbf{Q}_i\Lambda_0\mathbf{Q}_i' + \mathbf{R}^*_0 \quad (36)$$

Longitudinal Data and Covariance Functions (CF)

Covariance functions are another approach for dealing with longitudinal data. Meyer (1998) points out the similarity between covariance functions and the use of an RR model. A covariance function $U(x_1,x_2)$ is a function that describes the covariance between the measures of a randomly chosen individual at x_1 and the same individual at x_2 (Kirkpatrick and Heckman 1989, Kirkpatrick et al. 1990, Meyer and Hill 1997). Covariance functions were designed to deal with characters where the genetic effects can be expressed as a function dependent on continuous scales (for example, x_i is time or distance), like longitudinal data, morphological shape, and norms of reaction (Kirkpatrick and Heckman 1989). Thus, they are the continuous ("infinite-dimensional") equivalent to covariance matrices.

Kirkpatrick et al. (1990) presented a methodology using orthogonal polynomials to estimate covariance functions from a covariance matrix, later extended by Kirkpatrick et al. (1994). Essentially, the method has two steps. In the first step, a US model is used to estimate a covariance matrix. In the second step, the covariance function is truncated to the number of dimensions (or a reduced order) represented in the covariance matrix used to fit the function. If Φ is a matrix of orthogonal polynomials (Legendre polynomials in Kirkpatrick's work) with columns ϕ, $\mathbf{G}_0$ is a covariance matrix (e.g., additive genetic), and $\mathbf{U}_0$ is the covariance matrix of the polynomial coefficients then:

$$\mathbf{\Phi}_i = \begin{bmatrix} \phi_{01} & \phi_{11} & \cdots & \phi_{n1} \\ \phi_{02} & \phi_{12} & \cdots & \phi_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{0m} & \phi_{1m} & \cdots & \phi_{nm} \end{bmatrix} \tag{37}$$

$$\hat{U}(x_1,x_2) = \Sigma_i \Sigma_j \hat{\mathbf{U}}_{0ij}\, \phi_i(x_1)\phi_j(x_2) \tag{38}$$

where $\hat{\mathbf{U}}_0 = \Phi^{-1}\mathbf{G}_0\Phi'^{-1}$ for full fit, and it is estimated using generalized least squares when using reduced fit (see Kirkpatrick et al. 1990 for details). Full fit and reduced fit have the same meaning as in random regressions. Note that $\mathbf{Q}_i$ [Equation (26)] and Φ [Equation (37)] are equivalent if the same function is used to model the change of breeding values with time.

The estimation of covariance functions using Kirkpatrick's method relies on a previously estimated covariance matrix. Therefore it requires all individuals measured on a limited number of fixed ages, while a general specification of RR [as in Equation (23)] allows data spread over the trajectory without assumptions or restrictions for ages (van der Werf and Schaeffer 1997). Covariance functions permit interpolation and extrapolation of covariances in the same way as the RR model.

Considering the definition of covariance function [Equation (38) using $\hat{\mathbf{U}}_0 = \Phi^{-1}\mathbf{G}_0\Phi'^{-1}$], the RR model generates one of form $\mathbf{Q}_i\Lambda_0\mathbf{Q}_i'$. Nevertheless, the procedures are not identical. Although in RR fitting of a random effect depends on the fit of the other random effects [Equation (5) is solved for all variance components simultaneously], Kirkpatrick's method does not take into account other random effects (it considers only $\mathbf{G}_0$ and residuals are not "moved" into $\mathbf{R}_0$ to form $\mathbf{R}_0^*$). Other models partially provide the functionality of a covariance function. For example, the AR model (especially using a power formulation) can be used to span a correlation structure at any combination of times, but not to estimate the variances at each age, having then a more limited application.

Model Selection

A common approach to model selection is based on the likelihood ratio test (LRT), which asymptotically (i.e., with an "unspecified suitably large" number of observations), follows a chi-square distribution (Jones 1993). Two nested models (one model is a reduced version of the other), one with p independently adjusted parameters [rank($\mathbf{X}$) + number of covariance components] with log-likelihood $Log\ L_p$ and the other with $p + q$ parameters with log-likelihood $Log\ L_{p+q}$, are compared using:

$$LRT = 2(Log\ L_{p+q} - Log\ L_p) \sim \chi^2_q \tag{39}$$

The null hypothesis is that both models are the same (extra parameters do not improve the fit). Including more parameters in the model always increases or at least keeps the likelihood value; thus this test does not favor parsimonious models. There are several tests that take into account the number of parameters included in the model (see Jones 1993 for examples). One such test is Akaike's Information Criterion (AIC, Akaike 1974, Wada and Kashiwagi 1990), which is:

$$AIC = -2\ LogL + 2p \tag{40}$$

where $LogL$ is the log-likelihood and p the number of independently fitted parameters included in the model. The best model has the lowest value of AIC. If all models under comparison include the same fixed effects there is no need to consider rank ($\mathbf{X}$) in p, because it will not affect the differences on AIC.

Often the log-likelihood reported by statistical packages does not include the constant term [*con* in Equation (5)] because $LogL \propto LogL$ without *con*. Nevertheless, when comparing nonnested models (models with different distributional assumptions) the log-likelihood must use the complete density function, including all constants not involving the covariance parameters (Lindsey and Jones 1998).

Numerical Example

The use of different models is illustrated with basic wood density data (in kg/m^3) from breast-height increment cores of radiata pine (*Pinus radiata* D. Don) sampled from 28-yr-old open-pollinated families of the "268" series growing in Kaingaroa Forest, New Zealand (Shelbourne and Low 1980). The data set consists of 50 open-pollinated families with 5 blocks and 1 or 2 samples per block, i.e., families with 9 or 10 individuals totaling 424 trees. Each core contains between 20 and 28 measures of diameter at successive rings from the pith. Weighted basic density at age j (wbd_j) is calculated as:

$$wbd_j = \frac{\sum_{i=1}^{j} bd_i\Delta_i}{\sum_{i=1}^{j}\Delta_i} \tag{41}$$

where bd_i is the average basic density of ring i, and Δ_i is the area of ring i. Only weighted basic densities at ages 5, 10, 15, and 20 are considered in this example.

The general model utilized in the analyses is from Equations (6) to (8), where means per age are the only fixed effects. While some of the structures might not be

Table 1. Log-likelihood (LogL) and Akaike's information criterion (AIC) for the Unstructured (US), Full-fit Random Regressions (RRf), Banded Correlations (BC), Autoregressive (AR), Repeatability (RE), Uncorrelated (UC), and Reduced-fit Random Regressions (RRr) models.

Model	Parameters ($\mathbf{G}_0 + \mathbf{R}_0 = p$)	Log-likelihood ($LogL$)	AIC ($-2\ LogL + 2\ p$)
US and RRf	10 + 10	–4,886.90	9,813.80
BC	7 + 10	–4,891.03	9,816.06
AR	5 + 10	–4,892.71	9,815.42
RE	1 + 2	–6,327.04	12,660.08
UC	4 + 4	–6,715.32	13,446.64
RRr	6 + 10	–4,888.27	9,808.54

biologically plausible for a weighted density dataset (e.g., RE over a large number of years), we consider it appropriate to illustrate the effects of such models on the estimation of genetic parameters, and we include them in the analyses. All models are fitted using ASReml (Gilmour et al. 1998). Preliminary analyses considered blocks as random effects, but these were not significant and therefore excluded from subsequent models.

The log-likelihood ranged from –6715.02 for the UC model to –4886.90 for the US model, while AIC ranged from 9808.54 for the RR model to 13446.64 for the UC model (see Table 1). The AR and BC models have almost identical fitting but, considering AIC, the use of less parameters than in the US model reduced log-likelihood (Table 1). The RRr model was considered the most appropriate since it gave both the lowest AIC and estimates of genetic parameters closer to those of the US model (Table 2, Figure 2).

The scale effect is small, with phenotypic standard deviation ranging between 26.9 kg/m^3 (age 10) to 29.1 kg/m^3 (age 20). The data did not require transformation, as most models (except for RE) directly account for any heterogeneity of variances.

Heritabilities for age j (h_j^2) and genetic correlations between ages j and k (r_{jk}) were estimated with the following formulas, using corresponding elements from $\hat{\mathbf{G}}_0$ and $\hat{\mathbf{R}}_0$:

$$\hat{h}_j^2 = \frac{\hat{\sigma}^2_{a_j}}{\hat{\sigma}^2_{a_j} + \hat{\sigma}^2_{e_j}}$$

$$\hat{r}_{jk} = \frac{\hat{\sigma}_{a_{jk}}}{\hat{\sigma}_{a_j}\hat{\sigma}_{a_k}}$$

Table 2 presents genetic parameters estimates from the different models. As expected, the US and RRf (fitting a third-order orthogonal polynomial for each tree) models produce identical estimates of genetic parameters. The RRr model, which fits a second-order orthogonal polynomial for each tree, has a very similar fit with only six parameters in the $\mathbf{G}_0$ matrix.

In general, heritability estimates do not differ substantially among the models; however, the estimates for ages 5 and 20 are depressed in the AR and BC models, respectively (Table 2). This seems to be caused by the large reduction of the number of correlations estimated (especially with the AR model).

Table 2. Genetic parameters estimated from Unstructured (US), Full-fit Random Regressions (RRf), Banded Correlations (BC), Autoregressive (AR), Repeatability (RE), Uncorrelated (UC), and Reduced-fit Random Regressions (RRr) models. Heritability (h^2) and phenotypic variance (σ^2_p) and residual correlations (r_e, below diagonal).

			Age (yr)		
Age (yr)	h^2	σ^2_p	5	10	15
US and RRf					
5	0.731	792.411			
10	0.818	724.238	0.764		
15	0.805	782.797	0.537	0.837	
20	0.840	847.901	0.278	0.578	0.879
BC					
5	0.747	799.150			
10	0.823	725.560	0.595		
15	0.759	776.628	0.337	0.860	
20	0.771	837.059	0.118	0.677	0.918
AR					
5	0.678	792.432			
10	0.815	723.823	0.673		
15	0.786	780.333	0.330	0.821	
20	0.800	843.707	0.026	0.539	0.884
RE[a]					
5	0.567	1,052.468			
10	0.567	1,052.468	0.180		
15	0.567	1,052.468	0.180	0.180	
20	0.567	1,052.468	0.180	0.180	0.180
UC					
5	0.730	799.873			
10	0.818	726.869	0		
15	0.802	783.924	0	0	
20	0.815	847.186	0	0	0
RRr					
5	0.743	795.181			
10	0.802	721.986	0.713		
15	0.818	784.827	0.492	0.848	
20	0.842	848.130	0.210	0.607	0.869

[a] Heritability and phenotypic variance values apply across ages.

The results for the US and RRf additive genetic correlation structures are identical (Figure 2). The correlations between density at age 5 and later measurements are smaller than the correlations between successive measurements.

$$\mathbf{C}_{US} = \mathbf{C}_{RRf} = \begin{bmatrix} 1 & 0.941 & 0.881 & 0.846 \\ 0.941 & 1 & 0.987 & 0.968 \\ 0.881 & 0.987 & 1 & 0.993 \\ 0.846 & 0.968 & 0.993 & 1 \end{bmatrix}$$

The BC model constrains correlations with the same lag to be identical, estimating three correlations instead of six. Thence $\{0.941, 0.987, 0.993\} \rightarrow 0.988$, $\{0.881, 0.968\} \rightarrow 0.958$ and $\{0.846\} \rightarrow 0.917$ from the $\mathbf{C}_{US}$ (Figure 2). The banded BC model was not well suited to represent the correlations of age 5 with later measurements, overestimating the first column by values ranging from 0.047 to 0.077.

$$\mathbf{C}_{BC} = \begin{bmatrix} 1 & 0.988 & 0.958 & 0.917 \\ 0.988 & 1 & 0.988 & 0.958 \\ 0.958 & 0.988 & 1 & 0.988 \\ 0.917 & 0.958 & 0.988 & 1 \end{bmatrix}$$

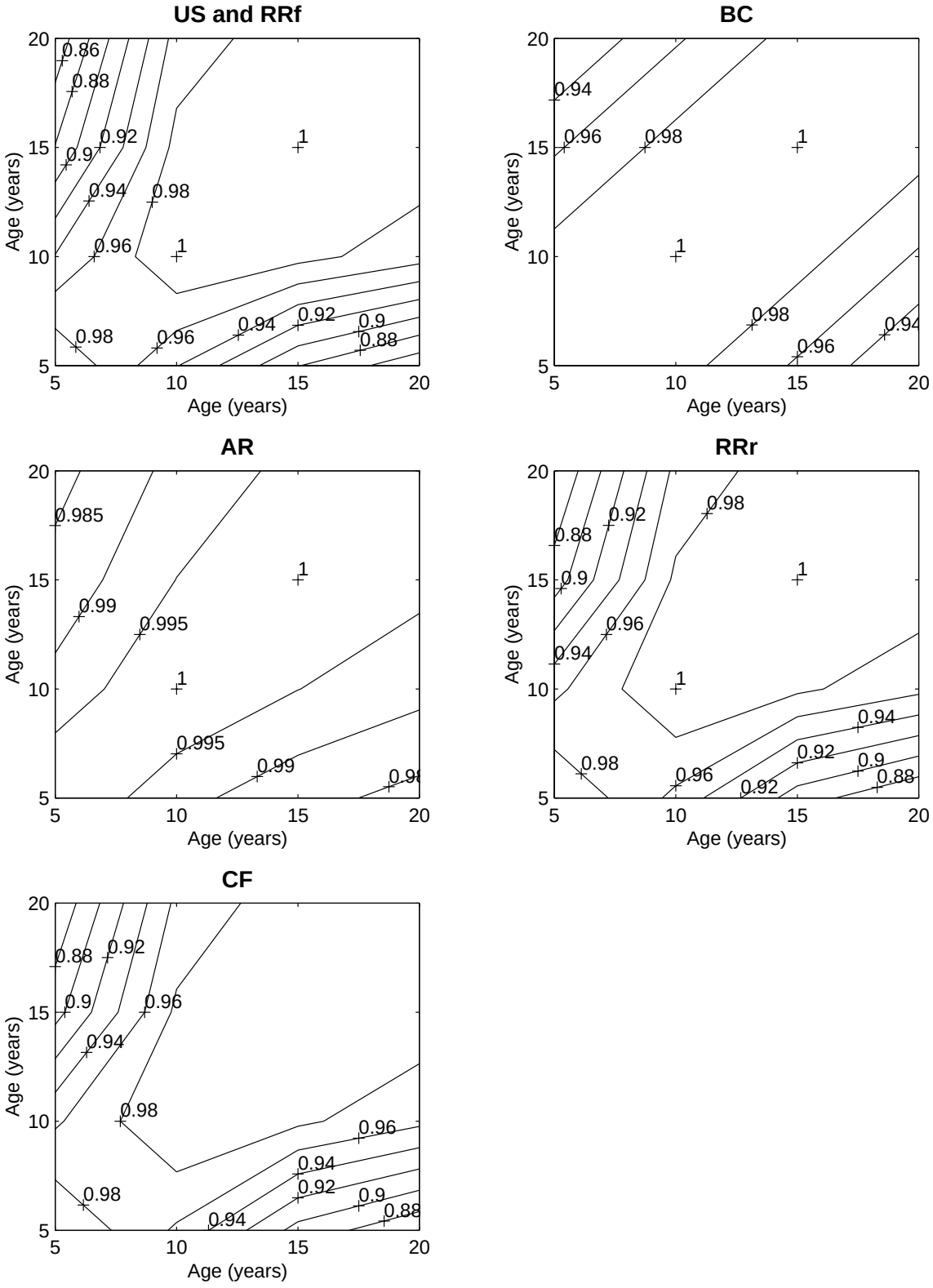

Figure 2. Contour plots of the correlation structures from the numerical example: US: unstructured, RRf: random regressions full fit (third degree polynomial), BC: banded correlations, AR: autoregressive, RRr: random regressions reduced fit (second degree polynomial), and CF: covariance function (second degree polynomial). Contour lines are labeled every 0.02 for all models except for the AR model, which is labeled every 0.005.

The AR model further reduces the number of parameters to be estimated. To achieve convergence it was necessary to use time in a natural logarithm scale, to accommodate ontogenetic effects. Thus the autocorrelation coefficient is expressed as $0.988^{|\log(age_k)-\log(age_j)|}$. Again the assumptions of the model are too restrictive, because a unique autoregression coefficient can not represent the lower correlation of the first measure with later ones. As a result all correlations are

overestimated. The spacing of the contour lines in Figure 2 was accordingly decreased from 0.020 to 0.005 for this model to improve presentation of results. The poor performance of the AR correlation matrix contrast with the results for tree height (m) obtained by Apiolaza et al. (2000) where it was selected as the best model.

$$\mathbf{C}_{AR} = \begin{bmatrix} 1 & 0.992 & 0.987 & 0.983 \\ 0.992 & 1 & 0.995 & 0.992 \\ 0.987 & 0.995 & 1 & 0.997 \\ 0.983 & 0.992 & 0.997 & 1 \end{bmatrix}$$

By definition the additive correlations are restricted to $\mathbf{C}_{RE} = \mathbf{J}$ [Equation (17)] and $\mathbf{C}_{UC} = \mathbf{I}$ [Equation (17)] for the RE and UC models, respectively. The RRr model (reduction from full-fit order 3 to order 2) appears to be less restrictive than the BC, AR, RE, and UC models and closely follows the results from the US model (Figure 2). This result also departs from the poor representation of genetic parameters for tree height reported by Apiolaza et al. (2000) for RRr models.

$$\mathbf{C}_{RRr} = \begin{bmatrix} 1 & 0.955 & 0.890 & 0.859 \\ 0.955 & 1 & 0.984 & 0.965 \\ 0.890 & 0.984 & 1 & 0.994 \\ 0.859 & 0.965 & 0.994 & 1 \end{bmatrix}$$

Residual correlation matrices of the US and RRr models were similar, as were the residual matrices of BC and AR (Table 2). Constraints in the UC and RE models rendered their residual correlation matrices distinct.

Results from covariance structures and covariance functions are not directly comparable, and we only present the additive genetic correlation matrix from the former approach. A covariance function, based on Legendre polynomials, is fitted to the $\mathbf{G}_0$ matrix from the US structure using a Mathematica notebook (Kirkpatrick et al. 1990).

$$\mathbf{C}_{CF} = \begin{bmatrix} 1 & 0.957 & 0.893 & 0.862 \\ 0.957 & 1 & 0.984 & 0.965 \\ 0.893 & 0.984 & 1 & 0.994 \\ 0.862 & 0.965 & 0.994 & 1 \end{bmatrix}$$

The results from the CF model are very similar to those from the US and RRf models, but require an estimate of the US structure as starting values. Again, fitting a second degree polynomial (i.e., six parameters for $\mathbf{G}_0$) appears to be an appropriate approximation to the results from the US model.

Final Remarks

The UC model has been applied in forestry, albeit implicitly, for studying changes of heritability with time. Covariances have typically been estimated by univariate analysis of the sums of pairs of measures, using the result $\text{Cov}(x,y) = [\text{Cov}(x + y) - \text{Var}(x) - \text{Var}(y)]/2$, but this does not allow unbiased use of data with missing observations such as occur from thinnings or mortality. The use of full multivariate evaluation takes into account the existence of selection or patterns of missing information; thus it provides unbiased minimum variance estimates of breeding values.

Breeders must be aware of large differences in the degree of parsimony, i.e., economy on the number of parameters to be estimated, and number and type of assumptions, involved in the different models presented. Hence, model selection should also consider biological plausibility of these assumptions. When there are only a few measurements, the US model (with no restricting assumption about the biological model) provides a good fit, but when increasing the number of measurements the probability of obtaining non-positive definite results increases. Using bending to obtain a positive definite matrix from the US model decreases the log-likelihood value, which may be lower than the ones coming from structured models (e.g., Apiolaza et al. 2000). The numerical example illustrates that it is necessary to find a compromise where the gains of using structures outweigh any bias due to model dependency. For example, the AR structure model involves the estimation of five parameters less than the US model, and reduces log-likelihood by only 5.8 units (for an AIC difference of 1.6) while providing a poor fit. On the other hand, the RR model requires four parameters less than the US model, reduces log-likelihood 1.4 units (with an AIC smaller by 5.3 units), and provides an almost perfect fit.

Different covariance structures have been compared in sheep breeding (Coelli et al. 1998 using US, BC, AR, and RE for fleece weight and fiber diameter) and tree breeding (Apiolaza et al. 2000 using US, BC, AR, RR, and UC for total height). These papers show that different traits need different models. Applications of RR are now popular in animal breeding, either using orthogonal polynomials (Meyer 1998, van der Werf et al. 1998), growth models (Jamrozik et al. 1997) or cubic splines (White et al. 1999). As pointed out by van der Werf et al. (1998), random regressions are an appealing approach, but in practice, covariance matrices estimated using the method can deviate significantly from those estimated using univariate or bivariate analyses. This behavior seems associated with strong reductions on the number of components (i.e., order of the polynomial compared to number of measures).

The fact that two models have similar AIC does not mean that their covariance matrices have similar "shape" (see Figure 2 and Apiolaza et al. 2000, as examples). Thus, while the objective is to reduce the number of parameters to be estimated, simultaneously the shape of the covariance matrices must be kept. Shaw (1991) suggests using maximum likelihood approach for the comparison of genetic covariance matrices, while Goodnight and Schwartz (1997) propose a bootstrap method.

Fitting multivariate models is certainly more complex and computationally demanding than using either a univariate approach (UC) or a series of bivariate analyses. On the other hand, it provides a description of the changes of genetic parameters with time. This article and Apiolaza et al. (2000) present both theory and examples for further optimization of

the breeding programs, considering number and timing of measurements of progeny tests, early selection, and an overall better understanding of the genetic control of traits subject to selection. Finally, it is necessary to point out that models of longitudinal data should consider any other effects present in the experiment (e.g., block, plots, etc.) in case they are relevant to the estimation of covariance components.

Literature Cited

AKAIKE, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19:716–723.

APIOLAZA, L.A., A.R. GILMOUR, AND D.J. GARRICK. 2000. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. Can. J. For. Res. 30:645–654.

BORRALHO, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. P. 141–145 *in* Proc. CRC-IUFRO conf. Eucalypts plantations: Improving fibre yield and quality, Potts, B.M., et al. (eds.). Hobart, Australia.

BURDON, R.D. 1989. Early selection in tree breeding: Principles for applying index selection and inferring input parameters. Can. J. For. Res. 19: 499–504.

CNAAN, A., N.M. LAIRD, AND P. SLASOR. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat. Med. 16: 2349–2380.

COELLI, K.A., A.R. GILMOUR, AND K.D. ATKINS. 1998. Comparison of genetic covariance models for annual measurements of fleece weight and fibre diameter. P. 31–34, Vol. 24 *in* Proc. 6th World Congress Genet. Appl. Livest. Prod. Armidale, Australia.

DIETERS, M.J., T.L. WHITE, R.C. LITELL, AND G.R. HODGE. 1995. Application of approximate variances of variance components and their ratios in genetic tests. Theor. Appl. Genet. 91:15–24.

DIGGLE, P.J. 1988. An approach to the analysis of repeated measurements. Biometrics 44: 959–971.

DIGGLE, P.J., K.-Y. LIANG, AND S.L. ZEGER. 1994. Analysis of longitudinal data. Clarendon Press, Oxford, UK. 253 p.

EVERITT, B.S. 1995. The analysis of repeated measures: a practical review with examples. Statistician 44:113–135.

FALCONER, D.S. 1952. The problem of environment and selection. Am. Natur. 86:293–298.

GABRIEL, K.R. 1962. Ante-dependence analysis of an ordered set of variables. Ann. Math. Stat. 33:201–212.

GILMOUR, A.R., R. THOMPSON, AND B.R. CULLIS. 1998. ASReml user's manual. New South Wales Agriculture, Orange, Australia. 170 p.

GOODNIGHT, J.H., AND J.M. SCHWARTZ. 1997. A bootstrap comparison of genetic covariance matrices. Biometrics 53:1026–1039.

GREGOIRE, T.G., O. SCHABENBERGER, AND J.P. BARRET. 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. Can. J. For. Res. 25:137–156.

HAND, D., AND M. CROWDER. 1996. Practical longitudinal data analysis. Chapman and Hall, London, UK. 232 p.

HENDERSON, C.R. 1984: Applications of linear models in animal breeding. University of Guelph Press, Guelph, Canada. 423 p.

HUBER, D.A., T.L. WHITE, AND G.R. HODGE. 1994. Variance component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. Theor. Appl. Genet. 88:236–242.

JAMROZIK, J., G.J. KISTEMAKER, J.C.M. DEKKERS, AND L.R. SCHAEFFER. 1997. Comparison of possible covariates for use in random regression model for analyses of test day yields. J. Dairy Sci. 80:2550–2556.

JENNRICH, R.I., AND M.D. SCHLUCHTER. 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42:802–820.

JONES, R.H. 1993. Longitudinal data with serial correlation: a state-space approach. Chapman & Hall, London, United Kingdom. 225 p.

KIRKPATRICK, M., AND N. HECKMAN. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. J. Math. Biol. 27:429–450.

KIRKPATRICK, M., D. LOFSVOLD, AND M. BULMER. 1990. Analysis of inheritance, selection and evolution of growth trajectories. Genet. 124:979–993.

KIRKPATRICK, M., W.G HILL, AND R. THOMPSON. 1994. Estimating the covariance structures for traits during growth and ageing, illustrated with lactation in dairy cattle. Genet. Res. 64:59–67.

LAIRD, N.M., AND J.H. WARE. 1982. Random-effects models for longitudinal data. Biometrics 38:963–974.

LINDSEY, J.K., AND B. JONES. 1998. Choosing among generalized linear models applied to medical data. Stat. Med. 17:59–68.

LOUIS, T.A. 1988. General methods for analysing repeated measures. Stat. Med. 7:29–45.

MAGNUSSEN, S., AND A. KREMER. 1993. Selection for an optimum tree growth curve. Silvae Genet. 42:322–335.

MEYER, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. Genet. Sel. Evol. 30:221–240.

MEYER, K., AND W.G. HILL. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or "repeated" records by restricted maximum likelihood. Livest. Prod. Sci. 47:185–200.

MRODE, R.A. 1996. Linear models for the prediction of animal breeding values. CAB International, Wallingford, UK. 187 p.

NETER, J., AND W. WASSERMAN. 1974. Applied linear statistical models. Richard D. Irwin, Homewood, IL. 842 p.

PATTERSON, H.D., AND R. THOMPSON. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554.

QUAAS, R.L., R.D. ANDERSON, AND A.R. GILMOUR. 1984. BLUP school handbook. Animal Genetics and Breeding Unit, University of New England, Australia. 158 p.

SEARLE, S.R. 1982. Matrix algebra useful for statistics. Wiley, New York. 438 p.

SHAW, R.G. 1991. The comparison of quantitative genetic parameters between populations. Evol. 45:143–151.

SHELBOURNE, C.J.A., AND C.B. LOW. 1980. Multi-trait index selection and associated genetic gains of *Pinus radiata* progenies at five sites. N. Z. J. For. Sci. 10:307–324.

SORIA, F., F. BASURCO, G. TOVAL, L. SILIÓ, M.C. RODRIGUEZ, AND M.A. TORO. 1997. Bayesian estimation of genetic parameters and provenance effects for height and diameter of *Eucalyptus globulus* in Spain. P. 95–100, Vol. 1 *in* Proc. IUFRO conf. on Silviculture and Improvement of Eucalypts. Salvador, Brazil.

VAN DER WERF, J.H.J., AND L. SCHAEFFER. 1997. Random regression in animal breeding. Course notes. CGIL, Guelph, Canada. 69 p.

VAN DER WERF, J.H.J., M.E. GODDARD, AND K. MEYER. 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. J. Dairy Sci. 81:3300–3308.

WADA, Y., AND N. KASHIWAGI. 1990. Selecting statistical models with information statistics. J. Dairy Sci. 73:3575–3582.

WEI, X., AND N.M.G. BORRALHO. 1998. Use of individual tree mixed models to account for mortality and selective thinning when estimating base population genetic parameters. For. Sci. 44:246–253.

WHITE, I.M.S., R. THOMPSON, AND S. BROTHERSTONE. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. J. Dairy Sci. 88:632–638.

APPENDIX 1. Direct Sum and Direct Product

The direct sum of n matrices $\mathbf{A}_i$ is defined as:

$$\Sigma_{\oplus} = \begin{bmatrix} \mathbf{A_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_n \end{bmatrix} = \text{diag}\{\mathbf{A}_i\} \qquad \text{(A1)}$$

Therefore, a direct sum of matrices creates a block diagonal matrix with the matrices being added in the diagonal and all off-diagonal elements equal to 0. Submatrices may be of different orders.

Example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \oplus \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{bmatrix}$$

The direct product of two matrices $\mathbf{A}_{pxq}$ and $\mathbf{B}_{mxn}$ creates a matrix where each submatrix is **B** multiplied by an element of **A**:

$$\mathbf{A}_{pxq} \otimes \mathbf{B}_{mxn} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & a_{pq}\mathbf{B} \end{bmatrix} \qquad \text{(A2)}$$

where a_{ij} is the element of **A** from row i and column j.

Example:

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \otimes \begin{bmatrix} 4 & 5 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 8 & 10 & 12 & 15 \\ 6 & 7 & 12 & 14 & 18 & 21 \end{bmatrix}$$