# MULTIPLE TRAIT IMPROVEMENT OF RADIATA PINE

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy in Forest Genetics and Breeding

at

Massey University

Palmerston North

New Zealand

Luis Alejandro Apiolaza

2000

## Abstract

This thesis explores the use of multivariate models in a tree breeding program with emphasis in radiata pine. It considers the development of breeding objectives, the efficiency of various strategies for subsampling trees to assess wood properties, and the analysis and exploitation of longitudinal data.

A model for a vertically integrated production system is developed — comprising forest production, pulp mill and sawmill — and evaluated for Chilean production and economic circumstances in each of three silvicultural regimes. The traits in the breeding objectives were volume at harvest age ($m^3$/ha) and average basic wood density ($kg/m^3$). Economic values for each trait were calculated as the difference in discounted profit for a unit marginal increase of volume or density. The objectives for different silvicultural regimes were similar, and a single objective — with relative weights 1:1.47 — appears to provide more economic gain than the use of specialist objectives.

Various subsampling schemes for wood property traits in progeny tests were studied through simulation in terms of reliability of estimates of genetic parameters, prediction of breeding values and expected genetic gains. Subsampling is subject to the Law of Diminishing returns, and measuring more than 15 trees per family did not provide large gains in accuracy of genetic parameters or in prediction of expected gain.

A unified view of multivariate analysis with longitudinal data from progeny trials is presented using a tree model. 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. Different models are compared for repeated assessments of basic wood density ($kg/m^3$). These models are further developed including additional random effects (block and plot) with an application to height (m) data using a Chilean radiata pine progeny test. Covariance structures reduce the risk of non-positive definite additive genetic matrices, while reducing computational demands for the analyses and providing a description of the genetic control of a trait over time.

Longitudinal data were used to predict breeding values close to rotation age, using either mass or combined selection. The method was tested under three covariance models and two breeding delays (time between selection and propagation of sufficient offspring for planting), to determine the best age — or combination of ages — for selection purposes. A combination of family information and repeated assessments provided the highest genetic gains.

*We do not receive wisdom, we must discover it by ourselves, after a journey through the wilderness, which no one else can make for us, which no one can spare us, for our wisdom is the point of view from which we come at last to regard the world.*

*—Marcel Proust*

*The journey is the destination.*

*—Anonymous*

*There is no intellectual exercise that is not ultimately useless.*

*—Jorge Luis Borges*

Dedicated to Marcela, for the love that sustains my days, Maite and Lucho for their support and love, and two precious souls, Haydée and Igor, who are always with me wherever I am.

## Acknowledgements

Dorian Garrick gave me both the ultimate blessing and curse: complete freedom to explore all topics and approaches I wanted. Although at times I used to think @#%& and ^%@#, I am really grateful for the learning experience. Mike Carson always provided useful general knowledge and background of tree breeding programs, as well as reassurance in difficult times. Thank you both for all the suggestions and advice.

I was very lucky to work with Rowland Burdon and Arthur Gilmour as co-authors in some of the chapters. Additionally, Rowland read and made comments for almost every chapter of the thesis. Both provided me with many opportunities to learn from their experience, new insights and numerous drafts to work on! Thank you both for your patience and sense of humour.

People at the New Zealand Forest Research Institute, Rotorua, provided help and comments at different stages of this work. Many thanks to Sue Carson (now a private consultant), Luis Gea, Keith Jayawickrama (now working in USA), Paul Jefferson, John Lee, Tony Shelbourne and Charles Sorensson (now with Fletcher Challenge). Special kudos goes to Luis Gea (a.k.a. Luigi), who helped me since before I arrived in New Zealand. He also provided me with lodging for every single trip to Rotorua. Many thanks for your help, advice, food, wine, etc. Thanks to John Lee for lending me his cycle workshop to sleep in.

My roommates provided a very supportive and fun environment for my work. It was great to share these years with Paul Charteris, Alastair Currie, Satish Kumar, Nicolás López-Villalobos, Fiona Miller, Richard Spelman and Claudia Ugarte. Many thanks to my friend Sarah Leberman and people at 'City Rock' for all the rock climbing together. Thanks to my friends on the phone and email: Cristián Estades (USA and Chile) and Ramy Alzamora (Chile). Many thanks to Frank Spirek and Diane Ensminger (Ogden, Utah, USA) for their friendship. They all helped me —maybe in vane— to keep my mental health while working on this thesis.

My PhD studies were funded by a New Zealand Official Development Assistance (NZODA) Scholarship and by a New Zealand Forest Research Institute (NZFRI)

Stipend. Margaret Smillie at Massey's International Student Office was an excellent scholarship administrator. I am grateful for her help during my studies.

My career has been influenced by the work of several people, who have been dealing with tree breeding programs since long before I started. The work done by Tim White, Tony Shelbourne, Rowland Burdon, Nuno Borralho and Steen Magnussen has been a constant source of inspiration. Tony Shelbourne is in part responsible for me being in New Zealand: thank you Tony.

And most of all, my infinite thanks to Marcela, who had more than her fair share of work during these last four years.

*Eres mi luna, eres mi sol,*
*mi dulze de mango, que dulzor.*
*—Alma Rosa*

**Post data**: Many thanks to Nuno Borralho (RAIZ, Portugal), Dorian Garrick (Massey University, New Zealand) and Dave Johnson (Livestock Improvement Corporation, New Zealand) for providing very positive feedback when reviewing the thesis. Allain Scott (IVABS, Massey University) helped me with the intricacies of submitting the final version of this thesis from overseas.

*May 2000*

## Table of contents

# CHAPTER ONE

# GENERAL INTRODUCTION

Tree breeding techniques have been applied for centuries, but industrial tree breeding programs started quite recently, within the last 50 years, e.g. Scots pine (*Pinus sylvestris*) and lodgepole pine (*Pinus contorta*) in Sweden, loblolly pine (*Pinus taeda*) and slash pine (*Pinus elliottii*) in United States, and radiata pine (*Pinus radiata*) in New Zealand (BANNISTER 1959, ZOBEL AND TALBERT 1984, WHITE et al. 1993, WILHELMSON and ANDERSSON 1995). Breeding programs involve multiple traits — normally a combination of form, growth rate and wood properties — that can be integrated in the selection process through independent culling and the use of selection indices. Even when interest is focussed on a single trait, forest operations cover multiple sites and/or multiple ages. Multiple traits or repeated expressions of a trait (in space or time) can be considered and fully exploited in a multivariate context.

Breeders are gradually orienting tree improvement programs towards end-products, rather than only forest-growth traits, acknowledging that most profit in the forest industry comes from the sale of elaborate products (SHELBOURNE, 1997). Breeding for end-products requires understanding of the relationships among tree-level traits and quantity/quality of end-products, as well as the economics of wood production and processing. Until five years ago, the definition of breeding objectives (in the sense of HAZEL 1943) in forestry was more the exception than the rule, and the few published objectives had a clear orientation towards pulp and paper production. Considering the end-product orientation and the existence of vertically integrated firms in the forestry sector, it is natural to include forest growth and several processing tiers (e.g. pulp mill and sawmill) when defining the model to construct the breeding objective (GREAVES 1999). Breeding objectives aim to maximise profit (often defined as a discounted cash flow), and profit is a complex function involving multiple traits.

Techniques applied to the evaluation of parents represented in progeny tests have experienced continuous technological advances, moving from estimating variance components using ANOVA methods and ranking trees using phenotypic family averages to combinations of restricted maximum likelihood (REML) and best linear unbiased prediction (BLUP) using tree ('animal') models (HENDERSON 1984, WHITE and HODGE 1989, BORRALHO 1995, HOFER 1998). Further sophistication is on the horizon, based on the use of multivariate methods to provide a detailed insight into changes in genetic control of traits over time. Longitudinal data arise when individuals

are assessed for the same outcome at repeated occasions during the lifetime of the individual. Covariance matrices of longitudinal data typically contain structures that can be modelled with a reduced number of parameters, improving estimation of covariance components (DIGGLE et al. 1994). Detailed analysis of longitudinal data may help to identify measurement and selection strategies that provide increased accuracy of prediction of breeding values at early ages (BURDON 1989).

The aim of this thesis is to explore the use of multivariate models in a tree breeding program with emphasis in, but not limited to, radiata pine. The main issues considered are the development of breeding objectives, sampling trees to assess wood properties, and the analysis and exploitation of longitudinal data.

Usually, breeding programmes assume that the selection objective is known, and concentrate on improving either the knowledge about the model equation (including the genetic parameters) or the selection strategy. Nevertheless, one of the weaknesses of tree improvement programs is the lack of formally defined breeding objectives. A breeding objective specifies an aggregate genotype where each trait is weighted by its own contribution to profit in the forest system. Chapter two asks the question 'what should we breed for?' and develops breeding objectives for three different silvicultural regimes of radiata pine in Chile. Although the number of assumptions is considerable, it is expected that simple objectives will provide an approximation close enough to the 'true' selection objective.

A typical breeding program considers both growth and form traits and properties of the resultant wood. The former are often cheap and easy to assess but with low heritabilities, while the latter are highly heritable but difficult and expensive to assess. For these reasons estimates of genetic parameters for wood properties (heritabilities and, especially, genetic correlations) are scarce and of unknown reliability, despite the importance of wood properties on the quality of end-products being well recognised (SORENSSON et al. 1997, SHELBOURNE 1997, EVANS et al. 1999). There is a need to use sampling schemes that provide good estimates of genetic parameters and breeding values per unit of cost. Chapter three reports research investigating the subsampling of wood properties while growth traits are measured in all trees. The effects on the

estimation of genetic parameters, prediction of breeding values and expected genetic gain are analysed through a simulation study.

Later chapters focus on changes in genetic parameters with time, viewed from a multivariate perspective. Chapter four provides a unified presentation of multivariate analyses focused on longitudinal data, with examples from a tree model perspective. Using the univariate tree model as foundation, repeated assessments are included and the concept of covariance structures is introduced. Covariance functions, an alternative approach used in evolutionary genetics (KIRKPATRICK and HECKMAN 1989), are also described. The relationship between alternative parameterisations is demonstrated. Chapter five further develops the multivariate models presented in the previous chapter, including additional random effects (block and plot) as well as providing an application to height data using a Chilean radiata pine progeny test.

Early selection is a recurrent topic in tree breeding, motivated by the long generation intervals that reduce gain per unit of time (e.g. KANG 1985, NEWMAN and WILLIAMS 1991, GWAZE et al. 1997). Chapter six presents the use of multiple assessments to predict breeding values close to rotation time, using either mass or combined selection. The method is tested under three covariance models and for two breeding delays (time between selection and propagation of sufficient offspring for planting), to determine the best age — or combination of ages — for selection purposes.

Chapter seven presents a general discussion of the thesis divided in three sections: breeding objectives, sampling progeny tests, and the use of longitudinal data and early selection. The discussion includes results from this thesis, relates them to results from previous research, and points out future directions of research. Finally the main conclusions of the thesis are presented.

## Literature cited


BANNISTER, M.H. 1959. Artificial selection and pinus radiata. *New Zealand Journal of Forestry* **8**: 69-90.

BORRALHO, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. P 141-145 *in* POTTS, B.M., BORRALHO, N.M.G., REID, J.B., CROMER, R.N., TIBBITS, W.N., and RAYMOND, C.A. "Eucalypts plantations: improving fibre yield and quality". Proceedings of CRC-IUFRO Conference, 19-24 February, Hobart, Tasmania, Australia.

EVANS, R., KIBBLEWHITE, R.P., LAUSBERG, M.J.F. 1999. Relationships between wood and pulp properties of twenty-five 13 year old radiata pine trees. *Appita Journal* **52**: 132-139.

GREAVES, B.L. 1999. Estimating an economic breeding objective for radiata pine grown for structural sawn-timber and liner-board. *Canadian Journal of Forest Research* (submitted).

GWAZE, D.P., WOOLLIAMS, J.A., and KANOWSKI, P.J. 1997. Optimum selection age for height in *Pinus taeda* L. in Zimbawe. *Silvae Genetica* **46**: 358-365.

HAZEL, L.N. 1943. The genetic basis for constructing selection indexes. *Genetics* **28**: 476-490.

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

HOFER, A. 1998. Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics* **115**: 247-265.

KANG, H. 1985. Juvenile selection in tree breeding: some mathematical models. *Silvae Genetics* **34**: 75-84.

KIRKPATRICK, M., and HECKMAN, N. 1989. A quantitative genetic model for growth, shape, reaction norm, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**: 429-450.

NEWMAN, D.H., and WILLIAMS, C.G. 1991. The incorporation of risk in optimal selection age determination. *Forest Science* **37**: 1350-1364.

SHELBOURNE, C.J.A. 1997. Genetics of adding value to the end-products of radiata pine. P 129-141 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

SORENSSON, C.T., COWN, D.J., RIDOUTT, B.G., TIAN, X. 1997. The significance of wood quality in tree breeding: a case study of radiata pine in New Zealand. P IV35-IV44 *in* CTIA/IUFRO International Wood Quality Workshop: Timber Management Towards Quality and End-Product Values. 18-22 August, Quebec City, Canada.

WHITE, T.L., and HODGE, G.R. 1989. Predicting breeding values with applications in forest tree improvement. Kluwer Academic Publishers, Boston.

WHITE, T.L., HODGE, G.R., and POWEL, G.L. 1993. An advanced-generation tree improvement plan for slash pine in the Southeastern United States. *Silvae Genetica* **42**: 359-371.

WILHELMSSON, L., and ANDERSSON, B. 1995. Breeding of Scots pine (*Pinus sylvestris*) and lodgepole pine (*Pinus contorta* ssp. *latifolia*). P 5-15 in Breeding programmes in Sweden. Arbetsrapport 302.

ZOBEL, B., and TALBERT, J. 1984. Applied forest tree improvement. John Wiley & Sons, New York.

# CHAPTER TWO

# BREEDING OBJECTIVES FOR THREE SILVICULTURAL REGIMES OF RADIATA PINE

Luis A. Apiolaza[1,2] and Dorian J. Garrick[1]

[1]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand. [2]New Zealand Forest Research Institute, Private Bag 3020, Rotorua, New Zealand.

SUBMITTED TO CANADIAN JOURNAL OF FOREST RESEARCH

## Abstract

A generic vertically integrated firm, comprising a production forest, a sawmill and a pulp mill was modelled under three silvicultural regimes: direct to pulp, intermediate (includes production thinning), and intensive (includes production thinnings and pruning). The harvest age traits included in the breeding objective were total volume ($m^3$/ha) and average wood density ($kg/m^3$). Economic values for each trait were calculated as the difference in discounted profit for a unit marginal increase of volume (Δvol) or density (Δden), and expressed as relative weights to facilitate comparisons between the objectives. The methodology was applied to a Chilean case study using representative economic and production circumstances. The breeding objectives so derived were 1 Δvol + 2.4 Δden for pulp, 1 Δvol + 1.1 Δden for intermediate, and 1 Δvol + 1.2 Δden for the intensive regime. The firm was profitable under all regimes. Genetic correlations between the objectives for each regime were higher than 0.9, indicating that a single breeding strategy with objective 1 Δvol + 1.5 Δden could be adopted, with almost no loss of genetic gain relative to selecting for a particular silvicultural regime.

## Introduction

The forest industry is a complex system where plantations are grown, harvested and then processed to produce many different end-products. Assuming that forest companies work in a competitive market, where the participants are interested in profit, it is possible to assume the industry is driven by profit maximisation. Profit, defined as the difference between discounted incomes and costs, is a complex function depending on many production and economic variables. An increase in profit can be achieved through increasing incomes (either higher quantity or quality of products), reducing costs or various combinations including both components. A better quality product should attract a price premium, while an incremental increase in volume of production at the same costs reduces average costs per unit.

Tree breeding is one of the tools utilised by the industry to increase profit. HAZEL (1943) formalised the concept of breeding objective (H) or aggregate economic genotype as a linear combination of additive genetic values of two or more traits weighted by their relative economic values:

$$H = v_1 a_1 + v_2 a_2 + \ldots + v_n a_n = \mathbf{v}' \mathbf{a}$$

where $\mathbf{v}' = [v_1\ v_2\ \ldots\ v_n]$ is the vector of relative economic values and $\mathbf{a}' = [a_1\ a_2\ \ldots\ a_n]$ is the vector of additive genetic values. An economic weight represents the benefit of one unit improvement of the trait without altering the other traits present in the objective (HAZEL 1943). Sometimes a distinction is made between absolute benefit (economic value) and relative benefit (economic weight) of improving a trait. The selection criteria in a breeding program are not necessarily the same as the traits in the objective, although their choice is dictated by the traits in the objective (BARLOW 1987, PONZONI and NEWMAN 1989). These criteria are normally combined in a selection index (I) with weights that maximise the correlation between H and I:

$$I = c_1 y_1 + c_2 y_2 + \ldots + c_m y_m = \mathbf{c}' \mathbf{y}$$

where $\mathbf{c}' = [c_1\ c_2\ \ldots\ c_m]$ is the vector of index weights calculated using genetic and economic information and $\mathbf{y}' = [y_1\ y_2\ \ldots\ y_m]$ is the selection criteria or vector of assessments (adjusted for fixed effects). If **G** is the m x n additive genetic covariance matrix among the m criteria in the index and the n traits in the objective, and **P** is the m x m phenotypic covariance matrix among the criteria in the index, $\mathbf{c} = \mathbf{P}^{-1}\ \mathbf{G}\ \mathbf{v}$. This index is optimal when genetic effects are completely additive and the economic weights are linear functions of their genetic value (GIBSON and KENNEDY 1990). Thus, knowledge of the breeding objective — traits and their economic weights — is a necessary condition to optimise selection in a breeding program incorporating multiple traits.

A good starting point to define a breeding objective is asking 'what do we want to improve?' (PONZONI and NEWMAN 1989). Usually it is possible to define the breeding objective as profit maximisation. The ideal breeding objective should comprise all traits which influence returns and costs, regardless of whether they can be measured or changed by selection (GJEDREM 1972, JAMES 1982, PONZONI 1982, BARLOW 1987, PONZONI and NEWMAN 1989). However, in practice the objective often only includes traits with reliable information and under genetic control (BORRALHO et al. 1993). Although in theory developing the breeding objective is the first step necessary to establish a breeding programme, this step has usually been postponed in tree breeding for several reasons:

1. Complexity of the forest processing industry.
2. Heterogeneity of wood properties, which make it difficult to ascertain relationships between wood properties and final quantity and quality of products.
3. A perception of breeding objectives research being of low priority.

While there have been previous attempts to value the contribution of tree breeding to industry profit (e.g. LÖFGREN 1988), BORRALHO et al. (1993) were the first to present a formal derivation of breeding objectives in forestry, using pulp production of *Eucalyptus globulus* in Portugal as an example. GREAVES and BORRALHO (1996) and GREAVES et al. (1997a) extended the model with a more complete description of the pulping system of eucalypts in Australia, while CHAMBERS et al. (1997) presented a breeding objective for thermo-mechanical pulping and newsprint production of radiata pine (*Pinus radiata*). LOWE et al. (1999) applied the general methodology of BORRALHO et al. (1993) to kraft and mechanical pulping of loblolly pine (*Pinus taeda*) in United States. All these models considered a cost minimisation perspective. SHELBOURNE (1997) emphasised the need to breed for added value to the end-product traits of radiata pine. SHELBOURNE et al. (1997) presented a comprehensive list of potentially important traits for different end-products of radiata pine. These traits might change both quantity and quality of end-products, or alter costs or incomes from industrial processes. GREAVES (1999) developed the first objective considering sawn-timber and accounting for quality of the product, expressed by structural grade.

The definition of a breeding objective requires study of the economic system that makes use of the trees available for breeding. The forest industry comprises several tiers (e.g. production forests, sawmills, pulp mills), and different products may have different requirements. Thus, the economic values are specific to the industry structure, and the nature of costs and incomes (e.g. see GROEN 1989 in an animal breeding context). At this stage it is necessary to define who benefits from the breeding program, because depending on the answer the objectives may differ (see AMER and FOX 1992 for a discussion). It is also necessary to define what are the restrictions (e.g. capital), if any, faced by the investors to define the profit criterion used in the breeding objective. With no restrictions net present value per ha may be an appropriate criterion; however, in the presence of capital constraints maximising the ratio of profit per unit of cost invested per ha (profitability index) may be more appropriate (ALLEN 1991).

Radiata pine is the main plantation species in Chile, generating more than 90% of the income from the forestry sector. Breeding of radiata pine in Chile started in the early 1980s, with the establishment of a university-industry co-operative program to serve the needs of a diversity of industrial processes. The objective of the research reported in this paper is to develop simple breeding objectives to maximise profit of a generic vertically integrated industry,

comprising a production forest, a sawmill and a pulp mill. Hence, the objective explicitly considers the value of end-products rather than assessing production values from a fixed return for raw materials. Three silvicultural regimes appropriate to sites of varying quality — pulp, intermediate and intensive — are considered, to represent part of the diversity of cost-income structure of the Chilean situation. Finally, we determine the compatibility of the breeding objectives for simultaneous application in a breeding program.

## Methods

### *Traits in the objective*

Two traits are considered in the breeding objective: harvestable volume ($m^3$/ha) and average wood density ($kg/m^3$), both expressed at rotation age (years). The choice of traits is based on their influence on profit, already demonstrated in previous studies (e.g. BORRALHO et al. 1993, GREAVES 1999), the possibility of constructing sensible relationships between forest and end-products (using either the objective traits or variables derived from them) and on the intent of keeping the breeding objective as simple as possible. A given volume per ha was described as a function of basal area ($m^2$/ha), top height (m) and stocking rate (trees/ha), accounting for individual variation of the trees that populate one hectare of forest.

### *Production system*

A vertically-integrated production system was modelled, which included tree growing, harvesting and processing in a sawmill and pulp mill (Figure 1), and considered three silvicultural regimes according to site quality expressed as site index (SI). Low quality sites (SI = 22) were managed with a pulp regime, medium quality sites (SI = 25) with an intermediate regime, and high quality sites (SI = 30) were under an intensive silvicultural regime. Table 1 provides a description of the characteristics of each regime.

**Figure 1:** Schematic diagram of a generic vertically integrated industry.

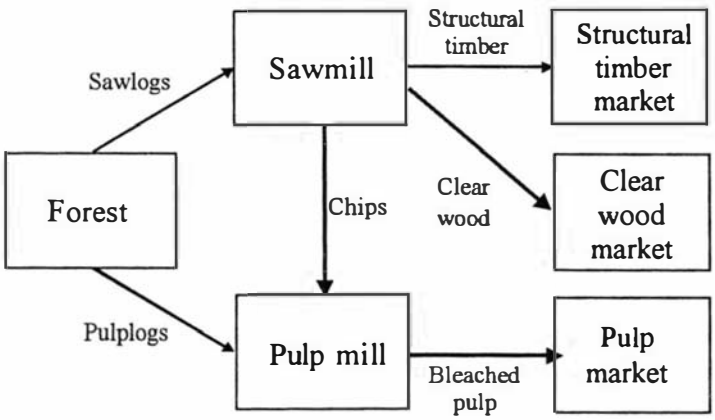



**Table 1:** Description of the silvicultural regimes.

| Descriptor | Silvicultural regime | | |
|---|---|---|---|
| | Pulp | Intermediate | Intensive |
| Site index (m)[a] | 22 | 25 | 30 |
| Rotation (years) | 18 | 22 | 26 |
| Initial stocking (stems/ha) | 1600 | 1400 | 1100 |
| Thinning to waste at age 5 (residual stems/ha) | 800 | 700 | 800 |
| Production thinning at age 12 (residual stems/ha) | - | 400 | 400 |
| Prunings (ages) | - | - | 5, 6, 7 |

[a] Average height of tallest 100 trees per hectare at age 20 years.

*Growth model*

A stand model based on a system of differential equations as described by GARCÍA (1984, 1994, 1999) was used to model the production forest tier, and provided expected values for basal area, top height and stocking rate. Average wood density was modelled using a quadratic model described by TIAN et al. (1995). Random sampling procedures were used to simulate the individual stems that contribute to the population of stems. Diameters of

individual trees were generated from stand parameters using a Weibull distribution, the parameters of which were obtained using a simplified method of moments (GARCÍA 1981). Total tree height was obtained using an exponential height-diameter function and wood density was derived using a normal distribution. Each silvicultural regime was simulated 100 times to explore the variability of results produced when generating individual trees using stochastic distributions.

*Log generation*

Trees were "cut" using three common log specifications based on length and small end diameter — [2.44 m, 10 cm], [3.60 m, 15 cm] and [4.00 m, 20 cm] respectively. Each stem was represented by 3 parameters: diameter at breast height, total height and wood density at breast height. The stem is compared with the largest log specification, if the remaining length and small end diameter — calculated through a taper function — exceeds the log specifications the tree is cut. Otherwise the model tries to fit the second largest log specification and so on, until the remainder of the tree is too small for any specification and is considered waste. This algorithm is a simplified version of that suggested by GOULDING and SHIRLEY (1979). An average wood density was assigned to each log using a parabolic function of height and wood density at breast height based on the model by TIAN et al. (1995).

*Log processing*

Logs of 2.44 m are used only for pulping. The weight of pulp for each log (TONpulp) was calculated as:

$$TONpulp = VOLlog * DENlog * PYlog$$

where volume of log calculation (VOLlog in $m^3$) is obtained from the large end diameter (LEDlog in cm), small end diameter (SEDlog in cm) and length (LENlog in m) of the log, and pulp yield (PYlog as proportion) using a function of basic wood density (DENlog) approximated from data in EVANS et al. (1999).

$$VOLlog = \pi * LENlog * (LEDlog^2 + LEDlog * SEDlog + SEDlog^2) /120000$$

$$PYlog = 0.37462 + 0.00024 *DENlog$$

The same function of yield is used for chips produced by the sawmill. Logs larger than 2.44 m are used in the sawmill, where the volume of timber is estimated as:

$$VOLtimber = LENlog * SEDlog^2 / 20000$$

In light framing construction, one of the main uses of radiata pine, stiffness (expressed as modulus of elasticity) is the most important property of the timber. Stiffness is closely related to wood density (COWN 1992). Modulus of elasticity of timber (MOEtimber in GPa) was estimated as a function of basic density (BIER 1985) and subsequently the grade of structural timber was obtained as a function of modulus of elasticity:

$$MOEtimber = -3.66 + 0.027 * DENlog$$

For logs 3.6 m long the volume of structural timber was:

$$VOLstruc = VOLtimber$$

In the case of 4 m long logs the proportion of timber that is clear wood is calculated using a second degree polynomial function based on data from Table 15 of COWN (1992), assuming a 20 cm diameter over stump (DOS):

$$VOLclear = VOLtimber * PROPclear$$

$$VOLstruc = VOLtimber * (1 - PROPclear)$$

$$PROPclear = -0.63309 + 0.0425 * SEDlog - 0.00032 * SEDlog^2$$

Ninety percent of the difference between volume of the log and volume of timber is considered chips while the residual is considered losses in the process (e.g. sawdust):

$$VOLchips = (VOLlog - VOLtimber) * 0.9$$

***Cost-income structure***

Discounted income (I) of the production system was derived from selling timber (structural and/or clear) and chemical pulp:

$$I = \sum_{k=1}^{r} (INCOMEstruc_k + INCOMEclear_k + INCOMEpulp_k)\, \lambda^k$$

where:

$r$ = rotation age

$\lambda = 1/(1 + \alpha)$

$\alpha$ = discount rate (cost of capital) expressed as proportion

$$INCOMEstruc_k = VOLstruc_k * PRICEstruc$$

$$INCOMEclear_k = VOLclear_k * PRICEclear$$

$$INCOMEpulp_k = TONpulp_k * PRICEpulp$$

Discounted cost (C) includes contributions from growing the forest, delivering the wood (roading, harvesting and transportation) and processing the logs. We assume there is no payment for raw materials between the processing and forest growing tiers (i.e. no transfer cost). Growth and delivery costs are summarised in Table 2. Hence:

$$C = \sum_{k=-1}^{r} (COSTgrowing_k + COSTdelivery_k + COSTprocessing_k)\ \lambda^k$$

where

$$COSTgrowing_k = COSTestab_k + COSTadmin_k + COSTinterv_k$$

**Table 2:** Cost structure for the silvicultural regimes.

| Item | Silvicultural regimes | | | | | |
|---|---|---|---|---|---|---|
| | Pulp | | Intermediate | | Intensive | |
| | US$ | Years | US$ | Years | US$ | Years |
| Cost of the land (per ha) | 640 | -1 | 750 | -1 | 900 | -1 |
| Preparation (per ha) | 70 | -1 | 70 | -1 | 70 | -1 |
| Establishment (per ha) | 180 | 0 | 170 | 0 | 160 | 0 |
| Weeding (per ha) | 50 | 1 | 50 | 1 | 50 | 1 |
| Administration (per ha) | 30 | annual | 30 | annual | 30 | annual |
| Waste thinning (per ha) | 60 | 5 | 50 | 5 | 50 | 5 |
| Roading (per ha) | 105 | [a] | | | | |
| Harvest cost (per $m^3$) | 7 | [b] | | | | |
| Transportation cost (per $m^3$) | 6 | [b c] | | | | |

[a] Roads are established for the first production intervention, i.e. final harvest (pulp) or production thinning (intermediate and intensive). See Table 1 for years.
[b] These costs are applied for thinnings and final harvest, and apply to all regimes.
[c] These costs assume an average hauling distance of 120 km.

$COSTestab_k$ considers cost of the land, preparation, planting and weeding, $COSTadmin_k$ are the annual administration costs, and $COSTinterv_k$ includes the costs of thinning to waste and prunings.

$$COSTdelivery_k = COSTroading_k + (VOLpulplogs_k + VOLsawlogs_k) * (COSTharvest + COSTtransport)$$

where $COSTroading_k > 0$ only for the first production intervention.

$$COSTprocessing_k = (VOLpulplogs_k + VOLchips_k) * COSTpulping +$$

$$VOLsawlogs_k * COSTsawing$$

where pulplogs include all 2.44m logs and sawlogs all 3.6m and 4m logs. The costs of processing a cubic meter of green wood in the sawmill and pulp mill (COSTsawing and COSTpulping respectively) were considered constant.

Discounted profit (P) or net present value is the difference between the discounted incomes and costs:

$$P = I - C$$

***Economic values***

Economics values were obtained from an incremental economic evaluation, where the value of a marginal increase of a trait was the difference between the actual cash flow and the base cash flow, i.e. with no increase on the trait (ALLEN 1991).

$$EVALUE = (P\Delta - Pbase) / (TRAIT\Delta - TRAITbase)$$

To simulate the increase in stand volume and average wood density rotations were extended between 1 and 2 years (depending on the silvicultural regime) to obtain stand parameters reflecting an increase of 10% on the traits. In this way changes to basal area, stocking and top height are accounted for. A 10% change in the traits was used because smaller changes are more difficult to reliably quantify in the model.

The prices of the end-products were: structural timber US$175/m$^3$, clear wood US$350/m$^3$ and bleached pulp US$430/ton. The sum of financial and operational costs (excluding cost of logs) for processing one cubic meter of green wood were US$35 in the pulp mill and US$45 in the sawmill. Discounted profit was calculated using a 10% ($\alpha = 0.1$) discount rate.

***Response to selection***

The response per generation in the objective H ($\Delta G_H$) when selecting on index I is (VAN VLECK et al 1987):

$$\Delta G_H = i\, r_{IH}\, \sigma_H$$

where i is the selection intensity, $r_{IH}$ is the correlation between the index and the objective (accuracy of prediction) and $\sigma_H$ is the standard deviation of the objective. Correlated response in objective $H_1$ ( $\Delta cG_{H_1}$ ) to selection for objective $H_2$ is calculated as:

$$\Delta cG_{H_1} = b_{H_1H_2}\, \Delta G_{H_2}$$

where $b_{H_1H_2}$ is the regression of $H_1$ on $H_2$. If both objectives have equal variance, this expression reduces to (see Appendix):

$$\Delta cG_{H_1} = r_{H_1H_2}\, \Delta G_{H_2}$$

where $r_{H_1H_2}$ is the genetic correlation between objectives $H_1$ and $H_2$. If the variances of the objectives are not equal the correlated response is only proportional to $r_{H_1H_2}$.

It is assumed that heritability for volume is 0.2, heritability for wood density is 0.6, and the genetic and phenotypic correlations between the traits are -0.3 and –0.1 respectively. Thus, additive genetic (**G**) and phenotypic (**P**) covariance matrices are:

$$\mathbf{G} = \begin{bmatrix} 350 & -141 \\ -141 & 630 \end{bmatrix} \text{ and } \mathbf{P} = \begin{bmatrix} 1750 & -135.55 \\ -135.55 & 1050 \end{bmatrix}$$

It is necessary to point out that genetic parameters involving wood properties have often been estimated with small sample sizes and/or inappropriate sampling schemes (see APIOLAZA et al. 1999), and therefore they are not completely reliable.

## Results and discussion

***Wood flow and cost-income structure***

Table 3 presents the average wood and end-product flow for 100 simulations of the base scenario for the silvicultural regimes. Total volume of logs (including production thinnings and final harvest) were 332.0 $m^3$/ha, 451.5 $m^3$/ha, and 794.0 $m^3$/ha for pulp, intermediate and intensive regimes respectively. Wood densities at harvest age were 405 kg/$m^3$, 404 kg/$m^3$ and 409 kg/$m^3$ respectively, reflecting differences in site quality (measured as site index) and rotation age.

The recovery rate for the sawmill (expressed as the ratio [structural volume + clear volume] / sawlogs volume) is 0.55 for intermediate and 0.56 for intensive regimes. Green wood requirements to produce a tonne of bleached pulp ([pulplogs + chips]/ bleached pulp) were, on average, 5.2 $m^3$, 5.4 $m^3$ and 4.9 $m^3$ for pulp, intermediate and intensive regimes respectively, showing the effect of age and site quality on basic wood density and, consequently, on pulp yield.

**Table 3:** Average wood and end-product flow per hectare for the production system under three silvicultural regimes (end-products in bold font).

| Products | Silvicultural regime | | | | |
|---|---|---|---|---|---|
| | Pulp | Intermediate | | Intensive | |
| | age = 18 | age = 12 | age = 22 | age = 12 | age = 26 |
| Pulplogs ($m^3$) | 332.0 | 42.3 | 72.3 | 119.3 | 65.5 |
| Sawlogs ($m^3$) | - | - | 336.9 | - | 609.2 |
| **Structural** ($m^3$) | - | - | 184.2 | - | 219.7 |
| **Clear wood** ($m^3$) | - | - | - | - | 123.1 |
| Chips ($m^3$) | - | - | 137.5 | - | 239.8 |
| **Bleached pulp** (ton) | 63.3 | 7.4 | 39.3 | 20.8 | 66.5 |

For the base scenario, as a proportion of total discounted costs, growing costs range between 24% (intensive) and 32% (pulp), delivery costs between 16% (intermediate) and 19% (pulp), and processing costs (pulping + sawing) between 49% (pulp) and 59% (intensive) (Table 4). In general, better sites and silviculture result in a shift in costs, as a percentage, towards the processing end. Concerning incomes, both intermediate and intensive were characterised by 60% in the sawmill and the remainder in the pulp mill (Table 4).

Increasing volume or density within a silvicultural regime did not noticeably change (more than 1%) the percentage cost structure reported in Table 4, but the percentages of income coming from pulp mill and sawmill were altered. In the intermediate regime sawing income starts at 56% for the base case and moves to 58% (Δvol) and to 53% (Δden). The percentages for the intensive regime move from 58% (base) to 61% (Δvol) and to 55% (Δden).

**Table 4:** Discounted cost and income structure expressed as percentage of total discounted costs and incomes for the base scenarios.

| Item | Silvicultural regime | | |
|---|---|---|---|
| | Pulp | Intermediate | Intensive |
| Costs | | | |
| Growing | 32 | 27 | 24 |
| Delivery [a] | 19 | 16 | 17 |
| Pulping | 49 | 25 | 30 |
| Sawing | - | 32 | 29 |
| Incomes | | | |
| Pulping [b] | 100 | 44 | 42 |
| Sawing | - | 56 | 58 |

[a] Includes production thinnings, final harvest, roading and transportation.
[b] Includes processing pulplogs and chips from sawlogs.

The base models for all regimes were profitable (Table 5), with discounted profit/ha (including forest, pulp mill and sawmill) ranging from US$659.06 (pulp) to US$4289.70 (intensive). The increase of harvested volume (due to higher site index and longer rotation) as well as the introduction of high value products (clear wood) justifies the difference in profit between pulp and more intensive regimes. Table 5 shows the change of profit by increasing ~10% final harvest volume (34 $m^3$/ha, 33 $m^3$/ha and 71 $m^3$/ha) and wood density (40.5 kg/$m^3$, 40.4 kg/$m^3$ and 40.9 kg/$m^3$) of the pulp, intermediate and intensive regimes. Percentage profit increase due to volume ranges between 18% (intermediate) and 32% (pulp), while the increase caused by density ranges between 14% (intensive) and 92% (pulp).

Economic values for each trait/regime combination (extra profit per unit increase) are presented in Table 5. The largest economic effect of increasing volume is in the intensive regime (11.94), reflecting the associated log size increase, improving sawmill recovery and quantity of structural and clear wood. Increasing density is more valuable in the pulp regime (14.90), because it directly improves the efficiency of pulp production. Thus the breeding objectives were: 6.26 $\Delta$vol + 14.90 $\Delta$den for pulp, 8.28 $\Delta$vol + 9.42 $\Delta$den for intermediate, and 11.94 $\Delta$vol + 14.70 $\Delta$den for the intensive regime. The similarities of economic values were more clearly interpreted when using relative weights. Relative economic weights for

volume and density were 1:2.4, 1:1.1, and 1:1.2 for pulp, intermediate and intensive regimes respectively.

**Table 5:** Discounted cash flow for the base, 10% volume increase (Δvol) and 10% density increase cases (Δden), where P is discounted profit and V is the economic value for a unit increase of volume (1 $m^3$) and wood density (1 $kg/m^3$).

| Model | Silvicultural regime | | | | | |
|---|---|---|---|---|---|---|
| | Pulp | | Intermediate | | Intensive | |
| | P | V | P | V | P | V |
| base | 659.06 | - | 1549.37 | - | 4289.70 | - |
| Δvol | 871.98 | 6.26 | 1822.57 | 8.28 | 5137.58 | 11.94 |
| Δden | 1262.34 | 14.90 | 1929.94 | 9.42 | 4890.91 | 14.70 |

***Multiple objectives***

The existence of several production conditions (e.g. silvicultural regimes) and economic circumstances (e.g. future end-product prices) creates sets of economic values for the traits under breeding. As a result breeders face the decision of keeping a single breeding population with a unique objective (either the objective for a specific condition or an average objective) or splitting the population and breeding for different objectives (DEL BOSQUE GONZÁLEZ and KINGHORN 1990, HOWARTH et al. 1997, HOWARTH and GODDARD 1998). The convenience of each option depends, among other factors, upon the correlation between the breeding values.

**Table 6:** Correlation between the breeding objectives, assuming heritability for volume is 0.2, heritability for density is 0.6 and additive genetic correlation between the traits is -0.3.

| Silvicultural regime | Pulp | Intermediate |
|---|---|---|
| Intermediate | 0.94 | |
| Intensive | 0.96 | 0.99 |

All correlations between breeding objectives reported in Table 6 are over 0.9. The variances of the breeding objectives were 127278.9, 57904.1, and 136537.9 for pulp, intermediate and intensive regimes respectively. Because the variances are different it is considered that correlated response is only proportional to the correlation between objectives. The objectives

for intermediate and intensive regimes are almost identical in the sense of ranking, although the economic values are different such the investment decisions may vary with regimes. Although splitting the breeding population in specialist 'breeds' would maximise individual response for each objective, this would simultaneously increase the breeding work and costs and reduce overall selection intensity. As an illustration, considering 600 plus trees with 30 progeny each, selection of top 200 for the next generation implies a selection intensity (i) of 2.634 (200/18000) for a single population, while the intensity is 2.231 (200/6000) when the population is split in three specialist 'breeds' (FALCONER and MACKAY 1996). The economic values for an 'average' breeding objective (see Appendix) are 8.83 and 13.01 ($[\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3] / 3$), i.e. the economic weights are 1:1.5. Direct responses to selection using specialist 'breeds' are US$ 589, US$ 349 and US$ 544 for pulp, intermediate and intensive regimes respectively, while direct response to selection on the average objective is US$ 579. Correlated response (see Appendix) on breeding objective for pulp, intermediate and intensive regimes to selection based on an average index will result in 1.10, 0.75 and 1.15 units of aggregate genotype per unit of response on the average objective. Considering these results and assuming other conditions identical — i.e. same selection criteria and generation interval — the firm is better off keeping a single breeding population.

## Final remarks

The profit in the production system may be utilised to determine the most appropriate payment system for raw materials, as well as to calculate the profit of the different tiers of the firm. There is a large degree of uncertainty in the cost-income structure of the generic firm, especially concerning future end-product prices. One option to tackle this problem is to change the discount rate used in the economic evaluation. Alternative approaches are to include the error probability distribution of the parameters estimated with uncertainty in the estimation of selection indices (AMER and HOFER 1994), or to run alternative realistic scenarios. Another problem is that part of the economic value of a trait can result from changes in the operation of the firm; thus operations of the firm must be optimised accordingly to reflect those changes (AMER et al. 1994). Most expected changes are in the silvicultural regimes (especially reduction of rotation age) rather than in the processing side of the firm. Because mill capacity often exceeds forest supply, increasing productivity per ha might encourage substitution of logs bought from third parties rather than changing mill size.

The breeding objectives modelled in this paper considered only two traits, currently included in the breeding strategy. There are several other traits with potential economic effect on the production system (e.g. SHELBOURNE et al. 1997). For example in sawmills recovery rate is affected by log sweep and taper (BROWN and MILLER 1975, KELLOGG and WARREN 1984). Mechanical performance of lumber (measured as stiffness and strength) is related to, apart from wood density, internode index and branch index (COWN 1992). Dimensional stability of planks depends on spiral grain. Lignin and cellulose content affect the yield of kraft pulp production, while the quality of pulp and paper is related to tracheid dimensions (KIBBLEWHITE et al. 1996). Moreover, there are other end-products and industrial processes (e.g. paper, fiberboards, veneers, carbon sequestration) that should be implemented in the model. Biological traits for some of these products have been listed elsewhere (SHELBOURNE et al. 1997). Once additional processing information is available to the breeder it is feasible to include additional traits in the objectives.

The calculation of economic values presented in this paper relies on the linearity of profit increase with changes on volume and density. This is not necessarily the case, as shown by GREAVES et al. (1997b) in eucalypts. Some traits are inherently non-linear, like those based on threshold values (e.g. spiral grain or grades for structural wood) where improvement of the trait has value only when reaching a given threshold (e.g COLLEAU and LE BIHAN-DUVAL 1995). There is a considerable body of research concerning selection on non-linear breeding objectives (e.g. GODDARD 1983, ITOH and YAMADA 1988, BURDON 1990), with varying conclusions concerning the importance of non-linearity. The relevance of non-linearity to practical selection and breeding programs is an issue that will need to be explored in detail.

## Acknowledgments

Luis Apiolaza was funded by NZODA and NZFRI scholarships. Many thanks to Ramy Alzamora (Instituto de Manejo Forestal, Universidad Austral de Chile), Verónica Alvarez and Ignacio Cerda (Departamento de Estudios Económicos, INFOR, Chile) for providing information regarding Chilean silvicultural regimes, prices and costs structures. Oscar García (University of Northern British Columbia) and Bob Shula (New Zealand Forest Research Institute) provided much help to understand growth models. The help of Bruce Greaves and Paul Chambers (University of Tasmania) providing unpublished material and valuable comments is much appreciated.

## Literature cited


ALLEN, D.H. 1991. Economic evaluation of projects- A guide. Stephen Austin & Sons, Hertford.

AMER, P.R., and FOX, G.C. 1992. Estimation of economic weights in genetic improvement using neoclassical production theory: an alternative to rescaling. *Animal Production* **54**: 341-350.

AMER, P.R., and HOFER, A. 1994. Optimum bias in selection index parameters estimated with uncertainty. *Journal of Animal Breeding and Genetics* **111**: 89-101.

AMER, P.R., FOX, G.C., and SMITH, C. 1994. Economic weights from profit equations: appraising their accuracy in the long run. *Animal production* **58**: 11-18.

APIOLAZA, L.A., BURDON, R.D., and GARRICK, D.J. 1999. Effect of univariate subsampling on the efficiency of bivariate parameter estimation and selection using half-sib progeny tests. *Forest Genetics* **6**: 79-87.

BARLOW, R. 1987. An introduction to breeding objectives for livestock. P 162-169 *in* Australian Association of Animal Breeding and Genetics. Proceedings of the Sixth Conference, 9-11 February, Perth, Western Australia, Australia.

BIER, H. 1985. Bending properties of structural timber from a 28-year-old stand of New Zealand *Pinus radiata*. *New Zealand Journal of Forestry Science* **15**: 233-250.

BORRALHO, N.M.G., COTTERILL, P.P., and KANOWSKI, P.J. 1993. Breeding objectives for pulp production of *Eucalyptus globulus* under different industrial costs structures. *Canadian Journal of Forest Research* **23**: 648-656.

BROWN, A.G., and MILLER, R.G. 1975. Effect of sweep on sawn recovery from radiata pine logs. *Australian Forest Research* **7**: 29-39.

BURDON, R.D. 1990. Implications of non-linear economic weights for breeding. *Theoretical and Applied Genetics* **79**: 65-71.

CHAMBERS, P.G.S., BORRALHO, N.M.G., BANHAM, P.W., and COX, R.E. 1997. Impact of wood selection traits on a thermo-mechanical pulping system using *Pinus radiata* to produce newsprint. P. 155-159 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

COLLEAU, J.J., and LE BIHAN-DUVAL, E. 1995. A simulation study of selection methods to improve mastitis resistance of dairy cows. *Journal of Dairy Science* **78**: 659-671.

COWN, D.J. 1992. New Zealand radiata pine and Douglas fir. Suitability for processing. FRI Bulletin 168.

DEL BOSQUE GONZÁLEZ, A.S., and KINGHORN, B.P. 1990. Implications of different selection objectives within open nucleus breeding schemes. P 95-102 *in* Australian Association of Animal Breeding and Genetics. Proceedings of the Eighth Conference, 5-9 February, Hamilton and Palmerston North, New Zealand.

EVANS, R., KIBBLEWHITE, R.P., and LAUSBERG, M. 1999. Relationships between wood and pulp properties of twenty-five 13 year old radiata pine trees. *Appita Journal* **52**: 132-139.

FALCONER, D.S., and MACKAY, T.F.C. 1996. Introduction to quantitative genetics. Longman, Essex.

GARCÍA, O. 1981. Note: Simplified method-of-moments estimation for the Weibull distribution. *New Zealand Journal of Forestry Science* **11**: 304-306.

GARCÍA, O. 1984. New class of growth models for even-aged stands: Pinus radiata in Golden Downs forest. *New Zealand Journal of Forestry Science* **14**: 65-88.

GARCÍA, O. 1994. The state-space approach in growth modelling. *Canadian Journal of Forest Research* **24**: 1894-1903.

GARCÍA, O. 1999. Height growth of Pinus radiata in New Zealand. *New Zealand Journal of Forestry Science* **29**: 131-145.

GIBSON, J.P., and KENNEDY, B.W. 1990. The use of constrained selection indexes in breeding for economic merit. *Theoretical and Applied Genetics* **80**: 801-805.

GJEDREM, T. 1972. A study of the definition of the aggregate genotype in a selection index. *Acta Agriculturae Scandinavica* **22**: 11-16.

GODDARD, M.E. 1983. Selection Indices for Non-linear Profit functions. *Theoretical and Applied Genetics* **64**: 339-344.

GOULDING, C.J., and SHIRLEY, J.W. 1979. A method to predict the yield of log assortments for long term planning. P 301-314 in ELLIOTT, D.A. (Comp.) "Mensuration for management planning of exotic forest plantations". NZ Forest Service, FRI Symposium No.20.

GREAVES, B.L. 1999. Estimating an economic breeding objective for radiata pine grown for structural sawn-timber and liner-board. *Canadian Journal of Forest Research* (submitted).

GREAVES, B.L., and BORRALHO, N.M.G. 1996. The influence of basic density and pulp yield on the cost of eucalypt kraft pulping: a theoretical model for tree breeding. *Appita Journal* **49**: 423-426.

GREAVES, B.L., BORRALHO, N.M.G., and RAYMOND, C.A. 1997a. Breeding objective for plantation eucalypts grown for production of kraft pulp. *Forest Science* **43**: 465-475.

GREAVES, B.L., BORRALHO, N.M.G., and RAYMOND, C.A. 1997b. Assumptions underlying the use of economic weights —are they valid in breeding for eucalypt pulp? *Forest Genetics* **4**: 35-42.

GROEN, A.F. 1989. Economic values in cattle breeding. I. Influences of production circumstances in situations without output limitations. *Livestock Production Science* **22**: 1-16.

HAZEL, L.N. 1943. The genetic basis for constructing selection indexes. *Genetics* **28**: 476-490.

HOWARTH, J.M., and GODDARD, M.E. 1998. Maximising response and profit under multiple objective selection. P 359-362 *in* Vol. 25 Proceedings of the 6th World Congress of Genetics Applied to Livestock Production, January 11-16, Armidale, New South Wales, Australia.

HOWARTH, J.M., GODDARD, M.E., and KINGHORN, B.P. 1997. Breeding strategies for targeting different breeding objectives. P 99-102 *in* Part One, Association for the Advancement of Animal Breeding and Genetics. Proceedings of the Twelfth Conference, 6-10 April, Dubbo, New South Wales, Australia.

ITOH, Y., and YAMADA, Y. 1988. Linear selection indices for non-linear profit functions. *Theoretical and Applied Genetics* **75**: 553-560.

JAMES, J.W. 1982. Economic aspects of developing breeding objectives: general considerations. P 107-117 *in* BARKER, J.S.F., HAMMOND, K., MCCLINTOCK, A.E. (Ed.). Future developments in the genetic improvement of animals, Academic Press, Australia.

KELLOGG, R.M., and WARREN, W.G. 1984. Evaluating Western hemlock stem characteristics in terms of lumber value. *Wood and Fiber Science* **16**: 583-597.

KIBBLEWHITE, R.P., EVANS, R., and RIDDELL, M.J.C. 1996. Handsheet property prediction from kraft fibre and wood tracheid properties in eleven radiata pine clones. Proceedings of the 50th Appita Annual General Conference, Auckland, New Zealand.

LÖFGREN, K.G. 1988. On the economic value of genetic progress in forestry. *Forest Science* **34**: 708-723.

LOWE, W.J., BYRAM, T.D., and BRIDGWATER, F.E. 1999. Selecting loblolly pine parents for seed orchards to minimize the cost of producing pulp. *Forest Science* **45**: 213-216.

PONZONI, R.W. 1982. Breeding objectives in sheep breeding programmes. P 619-634 Vol X *in* Proceedings of the 2nd World Congress of Genetics Applied to Livestock Production, Madrid, España.

PONZONI, R.W., and NEWMAN, S. 1989. Developing breeding objectives for Australian beef cattle production. *Animal Production* **49**: 35-47.

SEARLE, S.R. 1982. Matrix algebra useful for statistics. John Wiley & Sons, New York.

SHELBOURNE, C.J.A. 1997. Genetics of adding value to the end-products of radiata pine. P 129-141 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

SHELBOURNE, C.J.A., APIOLAZA, L.A., JAYAWICKRAMA, K.J.S., and SORENSSON, C.T. 1997. Developing breeding objectives for radiata pine in New Zealand. P 160-168 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

TIAN, X., COWN, D.J., and MCCONCHIE, D.L. 1995. Modelling of *Pinus radiata* wood properties. Part 2: basic density. *New Zealand Journal of Forestry Science* **25**: 214-230.

VAN VLECK, L.D., POLLAK, E.J., and OLTENACU, E.A.B. 1987. Genetics for the animal sciences. W.H. Freeman and Company, New York.

## Appendix: Some results on breeding objectives

### *Correlation between breeding objectives*

The correlation between two objectives $H_1$ and $H_2$ ( $r_{H_1H_2}$ ) is calculated as:

$$r_{H_1H_2} = \frac{Cov(H_1, H_2)}{\sqrt{Var(H_1)Var(H_2)}}$$

Variance and covariances of linear expressions are quadratic forms represented as a covariance matrix pre- and post-multiplied by a vector (SEARLE 1982, p 73). Hence:

$$r_{H_1H_2} = \mathbf{v}_1' \mathbf{G} \mathbf{v}_2 (\mathbf{v}_1' \mathbf{G} \mathbf{v}_1)^{-0.5} (\mathbf{v}_2' \mathbf{G} \mathbf{v}_2)^{-0.5}$$

### *Correlated response*

Correlated response in breeding objective 1 ($H_1$) when selection is based on an index derived to maximise response on breeding objective 2 ($H_2$) can be calculated using the regression of $H_1$ on $H_2$ ( $b_{H_1H_2}$ ):

$$\Delta cG_{H_1} = b_{H_1H_2} \Delta G_{H_2}$$

$$\Delta cG_{H_1} = \frac{Cov(H_1, H_2)}{Var(H_2)} \Delta G_{H_2} = \mathbf{v}_1' \mathbf{G} \mathbf{v}_2 (\mathbf{v}_2' \mathbf{G} \mathbf{v}_2)^{-1} \Delta G_{H_2}$$

Assuming that $Var(H_1) = Var(H_2)$, correlated response can be calculated as:

$$\Delta cG_{H_1} = \frac{Cov(H_1, H_2)}{Var(H_2)} \Delta G_{H_2} = \frac{Cov(H_1, H_2)}{\sqrt{Var(H_1)Var(H_2)}} \frac{\sqrt{Var(H_1)}}{\sqrt{Var(H_2)}} \Delta G_{H_2}$$

$$\Delta cG_{H_1} = r_{H_1H_2} \Delta G_{H_2}$$

which shows the relationship of response to the correlation between the objectives.

***Average breeding objective***

For each production condition or economic circumstance i there is a breeding objective $H_i = \mathbf{v}_i' \mathbf{a}$ and a selection index $\mathbf{c}_i = \mathbf{P}^{-1} \mathbf{G} \mathbf{v}_i$, with predicted genetic gain $\Delta G_i$. However, it might be possible to have only one generic breeding program, producing material for all sets of production and economic circumstances. The generic breeding objective is $H_g = \mathbf{v}_g' \mathbf{a}$ and a selection index $\mathbf{c}_g = \mathbf{P}^{-1} \mathbf{G} \mathbf{v}_g$, with predicted genetic gain $\Delta G_g$.

Considering t different breeding programs total gain using specialist breeding objectives is given by:

$$T\Delta G = w_1 \Delta G_1 + w_2 \Delta G_2 + \ldots + w_t \Delta G_t$$

where $w_i$ is a weight which includes the relative economic importance of the process and/or the plausibility of the economic circumstance. On the other hand, the total gain for using a generic breeding program is given by:

$$T\Delta G_g = w_1 \Delta cG_1 + w_2 \Delta cG_2 + \ldots + w_t \Delta cG_t$$

where $\Delta cG_i$ is the correlated response of breeding objective i when selecting for a generic breeding objective. Correlated response is calculated as:

$$\Delta cG_i = i\, \mathbf{c}_g` \mathbf{G} \mathbf{v}_i (\mathbf{c}_g` \mathbf{P} \mathbf{c}_g)^{-½} = i\, \mathbf{v}_g` \mathbf{G} \mathbf{P}^{-1} \mathbf{G} \mathbf{v}_i (\mathbf{v}_g` \mathbf{G} \mathbf{P}^{-1} \mathbf{G} \mathbf{v}_g)^{-½}$$

and considering $\mathbf{Q} = \mathbf{G} \mathbf{P}^{-1} \mathbf{G}$ then

$$\Delta cG_i = i\, \mathbf{v}_g` \mathbf{Q} \mathbf{v}_i (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-½}$$

The vector $\mathbf{v}_g$ should be calculated to minimise the difference between the total genetic gain produced by a generic breeding program and that obtained using several breeding programs or, the same result, maximise the total gain of the common breeding objective ($T\Delta G_g$):

$max\ T\Delta G_g = w_1 \Delta cG_1 + w_2 \Delta cG_2 + \ldots + w_t \Delta cG_t$

or equivalent expressions:

$max\ T\Delta G_g = \Sigma[w_i\, i\, \mathbf{v}_g` \mathbf{Q} \mathbf{v}_i (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-½}]$

$max\ T\Delta G_g = i\, \mathbf{v}_g` \mathbf{Q} \Sigma[w_i \mathbf{v}_i] (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-½}$

Differentiating $T\Delta G_g$ and setting the system of equations equal to 0:

$$\delta T\Delta G_g / \delta \mathbf{v}_g = i\, \mathbf{Q} \Sigma[w_i \mathbf{v}_i] (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-½} + i\, \mathbf{v}_g` \mathbf{Q} \Sigma[w_i \mathbf{v}_i] -½ (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-3/2} 2\, \mathbf{Q} \mathbf{v}_g = \mathbf{0}$$

$$i\, [\mathbf{Q} \Sigma[w_i \mathbf{v}_i] (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-½} - \mathbf{v}_g` \mathbf{Q} \Sigma[w_i \mathbf{v}_i] (\mathbf{v}_g` \mathbf{Q} \mathbf{v}_g)^{-3/2} \mathbf{Q} \mathbf{v}_g] = \mathbf{0}$$

$\mathbf{Q}\,\Sigma[w_i\,\mathbf{v}_i]\,(\mathbf{v}_g`\,\mathbf{Q}\,\mathbf{v}_g)^{-½} = \mathbf{v}_g`\,\mathbf{Q}\,\Sigma[w_i\,\mathbf{v}_i]\,(\mathbf{v}_g`\,\mathbf{Q}\,\mathbf{v}_g)^{-3/2}\,\mathbf{Q}\,\mathbf{v}_g$

$\mathbf{Q}\,\Sigma[w_i\,\mathbf{v}_i] = \mathbf{v}_g`\,\mathbf{Q}\,\Sigma[w_i\,\mathbf{v}_i]\,\mathbf{Q}\,\mathbf{v}_g\,(\mathbf{v}_g`\,\mathbf{Q}\,\mathbf{v}_g)^{-1}$

but $\mathbf{v}_g`\,\mathbf{Q}\,\Sigma[w_i\,\mathbf{v}_i]$ and $(\mathbf{v}_g`\,\mathbf{Q}\,\mathbf{v}_g)^{-1}$ are scalars, thus the solutions are proportional to their product (k), and $\mathbf{Q}$ is full rank so $\mathbf{Q}^{-1}$ exists.

$\mathbf{v}_g = k\,\Sigma[w_i\,\mathbf{v}_i]$

Therefore $\mathbf{v}_g$ is proportional to the weighted sum of vectors $\mathbf{v}_i$ (e.g. the weighted average). If all production conditions or economic circumstances are equally important, $\mathbf{v}_g$ is equal to the arithmetic mean of the $\mathbf{v}_i$.

# CHAPTER THREE

# EFFECT OF UNIVARIATE SUBSAMPLING ON THE EFFICIENCY OF BIVARIATE PARAMETER ESTIMATION AND SELECTION USING HALF-SIB PROGENY TESTS

Luis A. Apiolaza[1,2], Rowland D. Burdon[2] and Dorian J. Garrick[1]

[1] Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand. [2] New Zealand Forest Research Institute, Private Bag 3020, Rotorua, New Zealand.

FOREST GENETICS 6: 79-87 (1999)

## Abstract

Bivariate half-sib family data comprising 200 unrelated families with 30 individuals each were simulated for 80 combinations of genetic parameters (heritability for trait 1, heritability for trait 2 and genetic correlation between the traits) and random subsampling for trait 2 (3, 9, 15 and 30 trees). The model effects were all random, phenotypic variance of 1 for both traits and an environmental correlation of 0. The effect of subsampling was studied on: estimation of genetic parameters using restricted maximum likelihood (REML), best linear unbiased prediction (BLUP) of transmitting abilities, and expected response to selection. The lowest subsampling intensity generated greater biases, poorer representation of the distribution and larger coefficients of variation for estimates of genetic correlation and heritability of trait 2. The correlation between 'true' and predicted transmitting abilities for trait 2 had a direct relationship with subsampling intensity, heritability of the trait, and genetic correlation between traits 1 and 2. Even when the multivariate analysis increased the accuracy of prediction the correlation for trait 1 was only slightly affected. Direct response to index selection was depressed by low subsampling intensities, in a degree dependent on the heritabilities of the traits. Low subsampling boosted correlated responses for trait 1 and depressed those for trait 2. Truncation selection, subsampling trait 2 only in the top families for trait 1, was used with a specific set of parameters. This option produced the worst estimates and predictions. In summary, increasing the subsampled intensity gave progressively diminishing benefit, with little effect over 15 trees. A potential for improved cost-efficiency is thus confirmed.

## Introduction

Tree performance is typically a multi-trait function. For efficient genetic improvement reliable estimates of genetic parameters for the traits concerned are usually needed, to identify feasible breeding goals and to develop efficient selection procedures. These genetic parameters are encapsulated in the phenotypic (**P**) and genetic (**G**) variance-covariance matrices. Estimation of genetic parameters will always entail some form of population sampling. Often the sample will represent all trees in a progeny trial, but if a trait is very expensive to evaluate on individual trees subsampling is attractive, if not a necessity.

For selection, the cost efficiency of sampling of the available trees can be of twofold importance. In addition to affording the reliable genetic parameters estimates that are needed for constructing consistent selection indices, can allow reliable estimates of breeding values that may be needed for some traits.

Research in the last few years have confirmed the need for simultaneously considering growth traits and wood properties (BORRALHO et al. 1993, GREAVES and BORRALHO 1996, GREAVES et al. 1997, SHELBOURNE et al. 1997), making the issue of sampling more important. While wood properties may be important there is often a limited knowledge of their genetic parameters, especially the between-trait genetic correlations. Estimates of genetic parameters are usually obtained from analysis of progeny-test data using covariances among collateral relatives, e.g. half-sibs. While assessment of growth traits (usually of low heritability) in all the individuals of a progeny test is generally cheap and easy, satisfactory assessment of wood properties (usually highly heritable) is typically very costly per tree sampled. Accordingly it is usually appropriate to assess subsamples of relatively few individuals for wood properties, whereas many more individuals may be needed for providing good estimates of genetic parameters and breeding values for growth and form traits.

Several studies have focused on behaviour, in relation to sample size, of estimates of variance and covariance components, and thence of genetic correlation estimates. For example, ROBERTSON (1959), VAN VLECK and HENDERSON (1961), BROWN (1969), ROFF and PREZIOSI (1994) and LIU et al. (1997) have either described the distributions or given confidence intervals for estimated genetic parameters. Not explored was the issue of assessing, in the interest of cost-saving, a subsample of the study population for one of the variables. BURDON and APIOLAZA (1998) presented an ANOVA-based method to deal with two traits on partially overlapping subsamples, but it has some limits to the classification imbalance that can address. The objective of this study was to explore, through simulation, the effects of different subsampling intensities on the estimation of genetic parameters using restricted maximum likelihood (REML), on consistency of rankings based on Best Linear Unbiased Prediction (BLUP), and on estimates of expected response to selection.

## Materials and methods

The simulation experiment addressed the full factorial combinations of: heritability of trait 1 ($h_1^2$ = 0.1 and 0.3), heritability of trait 2 ($h_2^2$ = 0.4 and 0.8), genetic correlation between the traits ($r_g$ = -0.6, -0.3, 0, 0.3 and 0.6) and subsampling intensity of trait 2 (3, 9, 15 and 30 observations). Trait 1 was always considered with 30 trees (100% subsampling). One hundred progeny tests were simulated for each combination of levels of the factors. The tests were assumed for simplicity to have a completely random layout, 200 families and 30 individuals per family. Families were considered true half-sibs, with non-inbred, unrelated parents, and always a high number of effective paternal parents. Assuming a fully additive genetic model the coefficient of relationship was therefore ¼. Further assumptions were: 100% survival, phenotypic variance of 1 for both traits and an environmental correlation of 0. Even though the combinations of genetic parameters were not exhaustive, and that we used a fixed family number and size, we covered a range of situations that are relevant to tree breeding (see, for example, BURDON 1992, CORNELIUS 1994, WHITE 1987).

This study considered the prediction of transmitting abilities —½ of the breeding values— for backwards (or parental) selection. Therefore, for each progeny test of 200 models a family ('sire'[†]) model was fitted:

$$\mathbf{y} = \mathbf{X}\,\mathbf{m} + \mathbf{Z}\,\mathbf{f} + \mathbf{e}$$

where $\mathbf{y} = (\mathbf{y}_1' \; \mathbf{y}_2')'$ represents the vector of phenotypic observations for traits 1 and 2, $\mathbf{X} = \mathbf{X}_1 \oplus \mathbf{X}_2$ and $\mathbf{Z} = \mathbf{Z}_1 \oplus \mathbf{Z}_2$ are known incidence matrices for fixed and random effects respectively, $\mathbf{m} = (\mathbf{m}_1' \; \mathbf{m}_2')'$ and $\mathbf{f} = (\mathbf{f}_1' \; \mathbf{f}_2')'$ are vectors of unknown trait means and random family effects respectively, $\mathbf{e}$ is the vector of random residuals, ' is the transpose operation, and $\oplus$ is the direct sum operation. The expected value (E) and covariance (V) of the model equation terms are:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{e}\\ \mathbf{f}\end{bmatrix} = \begin{bmatrix}\mathbf{Xm}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \text{ and } V\begin{bmatrix}\mathbf{y}\\ \mathbf{e}\\ \mathbf{f}\end{bmatrix} = \begin{bmatrix}¼\mathbf{ZGZ}' + \mathbf{R} & \mathbf{R} & ¼\mathbf{ZG}\\ \mathbf{R} & \mathbf{R} & \mathbf{0}\\ ¼\mathbf{GZ}' & \mathbf{0} & ¼\mathbf{G}\end{bmatrix}$$

where:

$\mathbf{G} = \mathbf{G}_0 \otimes \mathbf{I}_N$ is the additive genetic variance-covariance matrix, where

[†] In tree breeding as opposed to animal breeding, the 'dam' and not the 'sire' is identified.

$\mathbf{G}_0 = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} \\ \sigma_{a_1 a_2} & \sigma_{a_2}^2 \end{bmatrix} = 4 \begin{bmatrix} \sigma_{f_1}^2 & \sigma_{f_1 f_2} \\ \sigma_{f_1 f_2} & \sigma_{f_2}^2 \end{bmatrix}$, $\otimes$ denotes the direct (Kronecker) product operation, and $\mathbf{I}_N$ is an identity matrix of order N equal to the total number of families (200). Because we are dealing with standardised traits, the phenotypic variance is 1 and as a result the genetic variances are $h_1^2$ and $h_2^2$ and the genetic covariances are $r_g\, h_1\, h_2$.
**R** is the residual variance-covariance matrix, which includes environmental effects and ¾ of the total genetic (co)variances. Because of the missing observations in trait 2, **R** cannot be expressed as a direct product. For individuals with records for both traits $\mathbf{R}_0 = \text{diag}\{\sigma_{e_1}^2, \sigma_{e_2}^2\}$; while for individuals with records only for trait 1, the matrix $\mathbf{R}_0$ collapses to the scalar $\sigma_{e_1}^2$.

Bivariate observations with the desired variance-covariance matrices were obtained using Cholesky decomposition (JOHNSON 1987, VAN VLECK 1994). Subsampling was accomplished by randomly deleting observations of trait 2 from the complete simulated test, leaving the desired number of trees in each family. Additionally, truncation subsampling was simulated for a specific set of parameters. In this case, all families were fully assessed and ranked for trait 1 and then 15 individuals from the top 40% of the families for that trait were randomly sampled for trait 2.

Variance and covariance components were estimated for each simulated test using REML (PATTERSON and THOMPSON 1971). An iterative average information algorithm was applied to maximise the likelihood function using AIREML (JOHNSON and THOMPSON 1995). Estimates of heritability ($h_i^2$) for traits 1 and 2, and genetic correlation between the traits ($r_g$) were calculated as:

$$\hat{h}_i^2 = \frac{4\hat{\sigma}_{f_i}^2}{(\hat{\sigma}_{f_i}^2 + \hat{\sigma}_{e_i}^2)}$$

$$\hat{r}_g = \frac{\hat{\sigma}_{f_1 f_2}}{\sqrt{\hat{\sigma}_{f_1}^2 \hat{\sigma}_{f_2}^2}}$$

with $\hat{\sigma}_{f_i}^2$ and $\hat{\sigma}_{f_i f_j}$ as the among-families estimated variance and covariance components respectively, and $\hat{\sigma}_{e_i}^2$ is the estimated variance of residuals.

For each simulated combination, the statistical significance of the skewness of distribution of estimates for a parameter i ($g_{\hat{p}_i}$, either for $\hat{h}_i^2$ or $\hat{r}_g$) was tested following SNEDECOR and COCHRAN (1980, p.78):

$$\frac{g_{\hat{p}_i}}{\left(\sigma_{\hat{p}_i}\right)^{\frac{3}{2}}} > crit_{\alpha,n}$$

where $\sigma_{\hat{p}_i}$ is the population standard deviation of the n values of $\hat{p}_i$ for a given combination of parameters, and $crit_{\alpha,n}$ is the tabulated critical value for a nominal probability of a comparisonwise type-I error, $\alpha$, with n degrees of freedom.

Bias for each simulated combination of parameters and sampling was considered significant if (LIU et al. 1997):

$$\frac{bias\sqrt{n}}{\hat{\sigma}_{bias}} > t_{\alpha,n-1}$$

where bias is the difference between the average of the n individual genetic parameter estimates and the 'true' (simulated) parameter, n is the number of simulations (100), $\hat{\sigma}_{bias}$ is the standard deviation of the n individual genetic parameter estimates, and $t_{\alpha, n-1}$ is Student's t value for a nominal probability of a comparisonwise type-I error $\alpha$ with n - 1 degrees of freedom.

Predicted transmitting abilities ($\hat{f}$) were obtained as solutions to Henderson's (1984) mixed model equations developed using the REML estimates of the variance components:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + 4\mathbf{G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{m}} \\ \hat{\mathbf{f}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

The effects of sampling on the prediction of transmitting abilities was quantified using the correlation ($r_{\hat{f}_i, f_{si}}$) between transmitting abilities for trait i predicted with a sample of the data ($\hat{f}_i$) and those actually simulated ($f_{si}$).

$$r_{\hat{f}_i, f_{si}} = \frac{\sigma_{\hat{f}_i, f_{si}}}{\sqrt{\sigma^2_{\hat{f}_i}\,\sigma^2_{f_{si}}}}$$

where $\sigma_{\hat{f}_i f_{si}}$, $\sigma^2_{\hat{f}_i}$ and $\sigma^2_{f_{si}}$ are the observed covariance and variance of predicted transmitting abilities using samples, and the variance of simulated transmitting abilities respectively.

The breeding objective (H) comprised a linear function of the transmitting abilities for traits 1 and 2 ($f_1$ and $f_2$). Selection was performed using an index (I), which included the bivariate-predicted transmitting abilities for traits 1 and 2 ($\hat{f}_1$ and $\hat{f}_2$). The relative economic values for traits 1 and 2 ($w_i$) were assumed in three separate situations as 1:1, 2:1 and 1:2. Thus,

$H = w_1 f_1 + w_2 f_2$

$I = w_1 \hat{f}_1 + w_2 \hat{f}_2$

The expected correlated ($\Delta_c G_i$, i.e. in the single trait i) and direct ($\Delta G_H$, i.e. in the breeding objective) responses to backwards selection on the index, i.e. selection of the parents based on progeny records, are given by (see Appendix):

$\Delta_c G_i = i\, \mathbf{w}' \mathbf{T}_i (\mathbf{w}' \mathbf{S} \mathbf{w})^{-1/2}$

$\Delta G_H = w_1 \Delta_c G_1 + w_2 \Delta_c G_2$

where $i$ is the selection intensity, $\mathbf{w}$ is the vector of relative economic weights, $\mathbf{T}_i$ is the vector of covariances between predicted transmitting abilities for both traits and true transmitting abilities for trait i, and $\mathbf{S}$ is the matrix of variances and covariances for predicted transmitting abilities.

## Results and discussion

### *Random Subsampling*

The results are presented separately for estimation of genetic parameters, prediction of transmitting abilities, and response to selection. The general trends are frequently exemplified using the parameters $h_1^2 = 0.1$, $h_2^2 = 0.8$ and $r_g = -0.3$, which can be applicable to *Pinus radiata* D. Don progeny tests assessing growth traits and wood properties. Most of the results presented for response to selection are based on relative economic weights of 2:1.

*Estimation of genetic parameters*

The different subsampling schemes were evaluated considering skewness, bias, and coefficient of variation of the estimates relative to the simulated 'true' parameters. The distributions of the REML estimates for heritabilities and genetic correlations were skewed, especially at low subsampling intensities and extreme heritabilities ($h_1^2 = 0.1$ or $h_2^2 = 0.8$). There, the likelihood approach constrained maxima to the parameter space tending to concentrate estimates close to the lower or upper bounds. Changes of sign for skewness while increasing subsampling intensities were commonplace. For example, skewness went from -0.49 to 0.42 for $\hat{h}_2^2$ and from 0.6 to -0.21 for $\hat{r}_g$, when subsampling 3 and 30 trees respectively (Table 1). Twenty-five out of 80 combinations of genetic parameters and subsampling presented significant skewness for $\alpha = 0.05$. The lowest subsampling intensities did not give a reliable representation of the distribution of the estimates.

The mean estimates of the genetic parameters varied slightly according to the different subsampling schemes. As a general trend, the magnitude of the bias of the estimates was higher for the lowest subsampling intensities (3 trees). The largest deviations were for $\hat{r}_g$, followed by $\hat{h}^2$ of traits 2 then 1. Twenty-four out of 80 combinations of genetic parameters and subsampling presented significant bias for $\alpha = 0.05$. Intensifying the subsampling from 3 to 15 trees reduced the bias (Table 1), but further intensification had little effect on the magnitude of bias.

The observed standard deviations of the estimates for each combination of parameters divided by the estimated means were considered as the 'empirical' coefficients of variation. As expected, an increased subsampling rate reduced the coefficient for all parameters estimates (Table 1). However, subsampling more than 9 trees gave only a marginal reduction in coefficient of variation.

**Table 1:** Bias, coefficient of variation (CV) and skewness (Skew) from 100 replicates using different subsampling intensities for trait 2, for $h_1^2 = 0.1$, $h_2^2 = 0.8$ and $r_g = -0.3$.

| Subsampling Intensity (number of trees) | $\hat{h}_1^2$ | | | $\hat{h}_2^2$ | | | $\hat{r}_g$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bias | CV | Skew | Bias | CV | Skew | Bias | CV | Skew |
| 3 | 0.001 | 0.020 | 0.246 | -0.016 | 0.020 | -0.497 | 0.014 | 0.050 | 0.601 |
| 9 | 0.000 | 0.020 | 0.249 | 0.008 | 0.012 | -0.225 | 0.013 | 0.043 | 0.323 |
| 15 | 0.000 | 0.020 | 0.253 | 0.006 | 0.010 | -0.127 | 0.005 | 0.036 | 0.055 |
| 30 | 0.000 | 0.020 | 0.253 | -0.002 | 0.008 | 0.421 | 0.005 | 0.036 | -0.216 |
| Tr[a] | 0.001 | 0.020 | 0.332 | -0.084 | 0.019 | -0.526 | 0.640 | 0.117 | -0.398 |

[a] Special case with truncation subsampling, using non-random selection of the assessed families.

Comparing the results of using coefficient of variation and percentile (data not shown), when the subsampling intensities were high, the trends for both standard error and percentile range using 15 trees were very close. However, the curves tended to differ when subsampling only 3 trees, showing the effects of the highly skewed distributions.

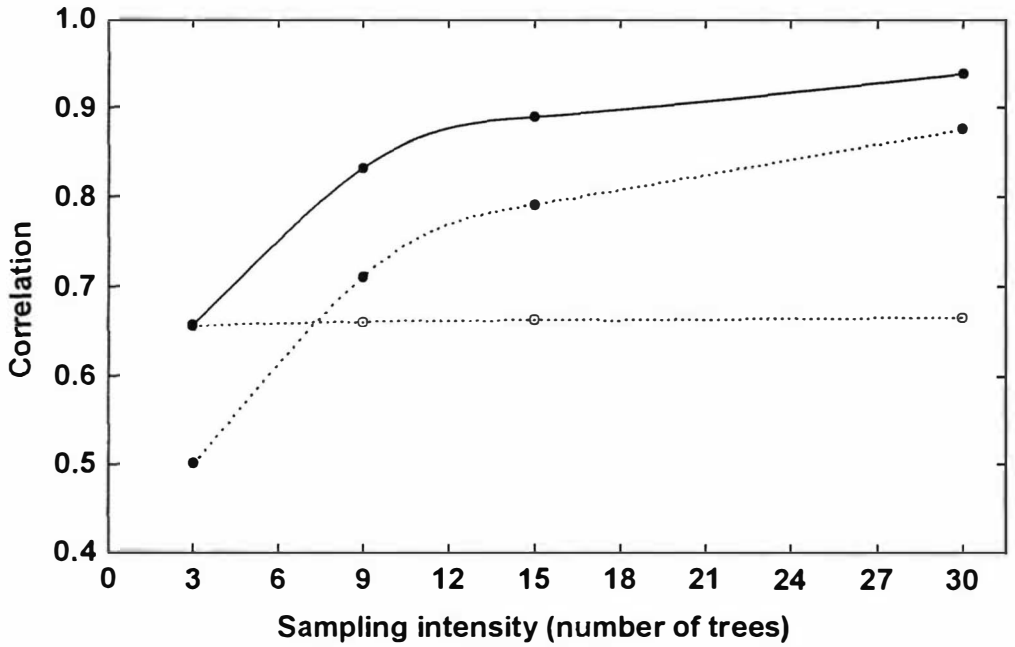
*Prediction of transmitting abilities*

A central part of the breeding process is the selection of the parents for the next generation. The effect of subsampling on the prediction of transmitting abilities was assessed using the correlation between transmitting abilities predicted for trait 1 ($\hat{f}_1$) and 2 ($\hat{f}_2$) using a subsample of the observations of trait 2 and those simulated ($f_{s1}$ and $f_{s2}$). The magnitude of the correlations involving trait 2 was strongly related to subsampling intensity (Figure 1, Table 2).

The correlation $r_{\hat{f}_2, f_{s2}}$ had a marked relationship with the heritability of the trait, where a higher $h_2^2$ was associated with a higher correlation. Thus, considering $h_1^2 = 0.1$ and $r_g = -0.3$, the correlations for $h_2^2 = 0.8$ exceeded those for $h_2^2 = 0.4$ by 0.15, 0.12, 0.09 and 0.06 for 3, 9, 15 and 30 subsampled trees respectively (Figure 1). Simultaneously, a higher simulated correlation $r_g$ between the traits (either positive or negative) gave higher correlations with predicted transmitting abilities (Table 2). This difference was

noteworthy for the lowest subsampling intensity, with magnitude of up to 0.15 (for $h_1^2$ = 0.3, $h_2^2$ = 0.4, 3 subsampled trees, and $r_g$=0 and 0.6). The effect for the correlation was symmetric; that is, an increase of $r_g$ in either way produced essentially the same increase in $r_{\hat{f}_2,f_{s2}}$. The decrease in heritability of trait 1 tended to accentuate the effect of different $r_g$ on $r_{\hat{f}_2,f_{s2}}$ (Table 2).

**Figure 1:** Correlation between transmitting abilities predicted using 3, 9, 15 and 30 observations and those simulated, considering $r_{\hat{f}_1,f_{s1}}$ for $h_1^2$ = 0.1, $h_2^2$ = 0.4 and $r_g$ = -0.3 (···o···), and $r_{\hat{f}_2,f_{s2}}$ for $h_1^2$ = 0.1, $r_g$ = -0.3 and $h_2^2$ = 0.4 (···•···) or $h_2^2$ = 0.8 (–•–).




Even though trait 1 was not subject to subsampling, the correlations $r_{\hat{f}_1,f_{s1}}$ still rose slightly with increased subsampling for trait 2, because the multivariate analysis increased the accuracy of prediction when including information from trait 2 (THOMPSON and MEYER 1986). However, the effect of subsampling on trait 2 was small compared with the results for $r_{\hat{f}_2,f_{s2}}$ (Figure 1, Table 2). The effect of including trait 2 depended on the value of $r_g$; thus the highest $r_{\hat{f}_1,f_{s1}}$ were for high $r_g$ (Table 2). In other words, a strong association between the traits contributed to more reliable rankings of parents for trait 1 when only subsampling trait 2. In general, across all subsampling intensities and

parameter combinations, $r_{\hat{f}_1,f_{s1}}$ ranged from 0.65 to 0.85, indicating a considerable agreement between the selection of parents under the different subsampling intensities.

**Table 2:** Correlation between transmitting abilities predicted using 3, 9, 15 and 30 observations and those simulated, considering all the combinations of genetic parameters, for $r_{\hat{f}_1,f_{s1}}$ and $r_{\hat{f}_2,f_{s2}}$.

| Genetic parameters | | | Subsampling intensity (number of trees) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 3 | | 9 | | 15 | | 30 | |
| $h_1^2$ | $h_2^2$ | $r_g$ | $r_{\hat{f}_1,f_{s1}}$ | $r_{\hat{f}_2,f_{s2}}$ | $r_{\hat{f}_1,f_{s1}}$ | $r_{\hat{f}_2,f_{s2}}$ | $r_{\hat{f}_1,f_{s1}}$ | $r_{\hat{f}_2,f_{s2}}$ | $r_{\hat{f}_1,f_{s1}}$ | $r_{\hat{f}_2,f_{s2}}$ |
| 0.1 | 0.4 | -0.6 | 0.67 | 0.55 | 0.69 | 0.72 | 0.70 | 0.80 | 0.71 | 0.88 |
| | | -0.3 | 0.66 | 0.50 | 0.66 | 0.71 | 0.66 | 0.79 | 0.67 | 0.88 |
| | | 0 | 0.66 | 0.48 | 0.66 | 0.70 | 0.66 | 0.79 | 0.66 | 0.88 |
| | | 0.3 | 0.65 | 0.49 | 0.66 | 0.71 | 0.66 | 0.79 | 0.67 | 0.87 |
| | | 0.6 | 0.67 | 0.57 | 0.69 | 0.73 | 0.70 | 0.79 | 0.72 | 0.88 |
| | 0.8 | -0.6 | 0.65 | 0.68 | 0.70 | 0.84 | 0.72 | 0.89 | 0.73 | 0.94 |
| | | -0.3 | 0.65 | 0.66 | 0.68 | 0.83 | 0.67 | 0.89 | 0.67 | 0.94 |
| | | 0 | 0.60 | 0.65 | 0.65 | 0.83 | 0.65 | 0.89 | 0.65 | 0.94 |
| | | 0.3 | 0.65 | 0.65 | 0.67 | 0.83 | 0.67 | 0.89 | 0.67 | 0.94 |
| | | 0.6 | 0.68 | 0.68 | 0.71 | 0.84 | 0.72 | 0.89 | 0.72 | 0.94 |
| 0.3 | 0.4 | -0.6 | 0.84 | 0.61 | 0.85 | 0.75 | 0.85 | 0.81 | 0.85 | 0.88 |
| | | -0.3 | 0.84 | 0.51 | 0.84 | 0.71 | 0.84 | 0.79 | 0.85 | 0.88 |
| | | 0 | 0.84 | 0.46 | 0.84 | 0.70 | 0.84 | 0.78 | 0.84 | 0.87 |
| | | 0.3 | 0.84 | 0.52 | 0.84 | 0.72 | 0.84 | 0.80 | 0.84 | 0.88 |
| | | 0.6 | 0.84 | 0.61 | 0.85 | 0.75 | 0.85 | 0.81 | 0.85 | 0.88 |
| | 0.8 | -0.6 | 0.82 | 0.70 | 0.85 | 0.84 | 0.85 | 0.89 | 0.85 | 0.94 |
| | | -0.3 | 0.79 | 0.66 | 0.83 | 0.83 | 0.84 | 0.89 | 0.84 | 0.94 |
| | | 0 | 0.76 | 0.65 | 0.82 | 0.83 | 0.84 | 0.89 | 0.84 | 0.94 |
| | | 0.3 | 0.79 | 0.66 | 0.82 | 0.84 | 0.84 | 0.89 | 0.84 | 0.94 |
| | | 0.6 | 0.85 | 0.71 | 0.85 | 0.85 | 0.85 | 0.89 | 0.85 | 0.94 |

*Response to selection*

To study the effect of subsampling, direct and correlated responses are presented for each combination of genetic parameters as the ratio of response of a given subsampling intensity (using 3, 9, 15 and 30 trees) over the response predicted using the real parameters.

The effects of subsampling were evident in the estimation of direct ($\Delta G_H$) and correlated ($\Delta_c G_i$) responses. Low subsampling intensities consistently depressed the predicted response compared to the case with true genetic parameters (Figure 2). Direct response was very dependent on the heritabilities of the traits (Figure 2) and, for the lower subsampling intensities, on the genetic correlation between them (Table 3). The effect of genetic correlation was dependent on its magnitude and sign. Increasing subsampling reduces the range of relative direct gain between $r_g = -0.6$ and $r_g = 0.6$. For example, for $h_1^2 = 0.1$ and $h_2^2 = 0.8$ the range is 0.25, 0.20, 0.13 and 0.01 for 3, 9, 15 and 30 subsampled trees respectively (Table 3).

For all the genetic parameter combinations, the average direct response achieved 90% of the expected response subsampling just 15 trees (Figure 2). Nevertheless, there can be lower values for combinations of low heritabilities and negative genetic correlations (e.g. for $h_1^2 = 0.1$, $h_2^2 = 0.8$ and $r_g = -0.6$, Table 3). The effect of increasing subsampling in trait 2 reduced the difference in average response to selection among combinations of heritability. As an example, the difference between $h_1^2 = 0.1$, $h_2^2 = 0.4$ and $h_1^2 = 0.3$, $h_2^2 = 0.4$ changes from 0.18 to 0.05 subsampling 3 and 15 trees respectively (Figure 2). With further subsampling the difference converges to 0, reflecting diminishing cost-efficiency of additional sampling of trait 2.

**Figure 2:** Average relative gain for different subsampling intensities, expressed as the ratio of response to the predicted direct response using the true parameters. Results presented for relative economic weights 2 (trait 1) and 1 (trait 2), considering $h_1^2 = 0.1$ and $h_2^2 = 0.4$ (···o···), $h_1^2 = 0.1$ and $h_2^2 = 0.8$ (···•···), $h_1^2 = 0.3$ and $h_2^2 = 0.4$ (–o–), and $h_1^2 = 0.3$ and $h_2^2 = 0.8$ (–•–).

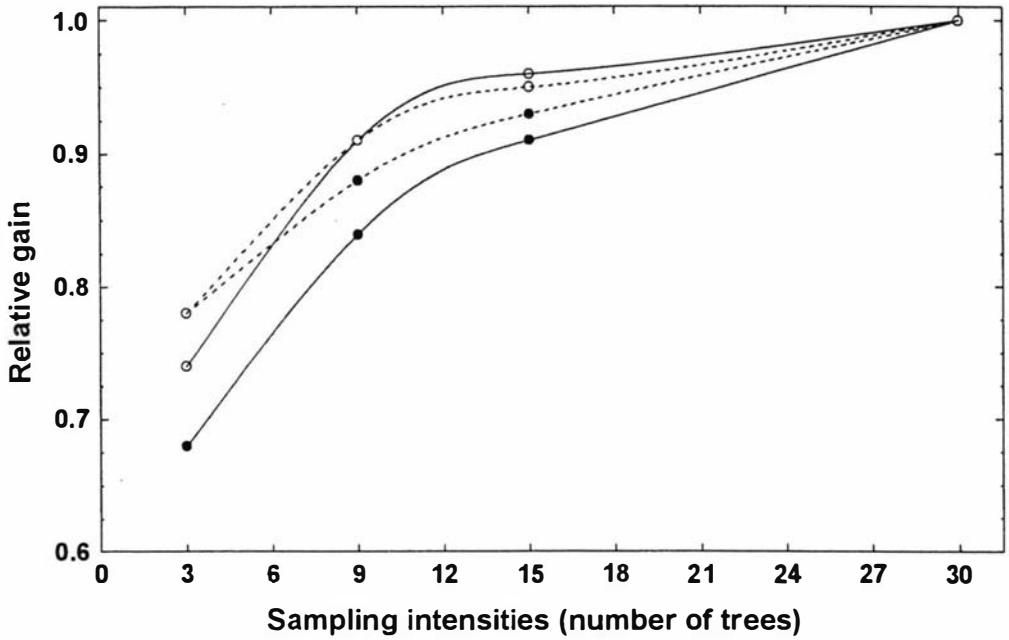


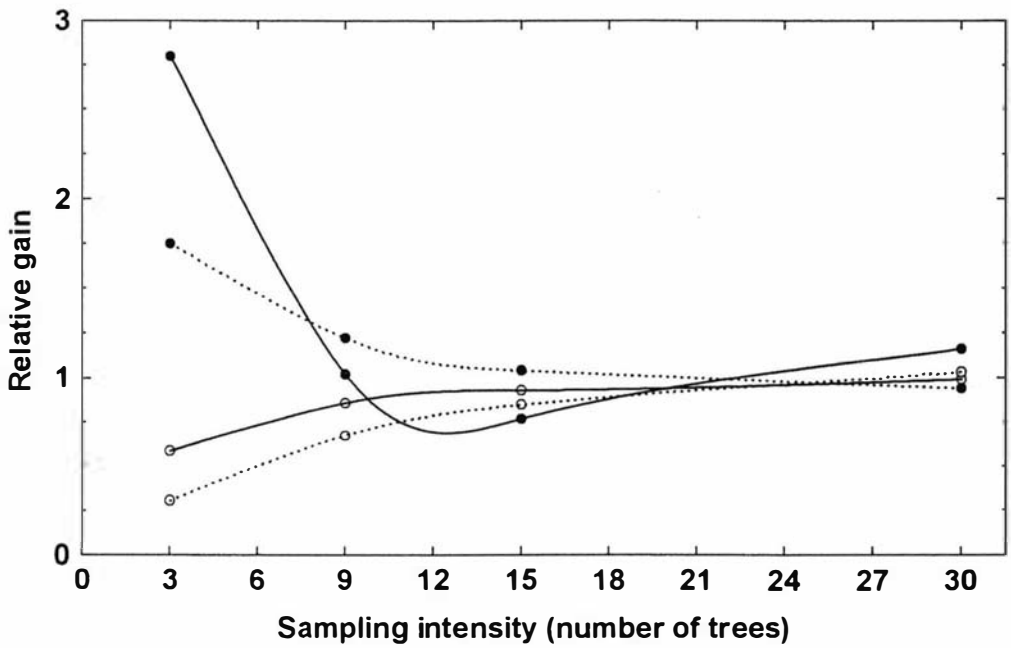

**Table 3:** Effect of subsampling on the relative direct response to selection for economic weights 2:1, $h_1^2 = 0.1$, $h_2^2 = 0.8$ and a range of genetic correlations between the traits.

| Genetic correlation | Subsampling intensity (number of trees) | | | |
|---|---|---|---|---|
| | 3 | 9 | 15 | 30 |
| -0.6 | 0.64 | 0.80 | 0.87 | 0.99 |
| -0.3 | 0.70 | 0.87 | 0.92 | 1.00 |
| 0 | 0.75 | 0.90 | 0.95 | 0.99 |
| 0.3 | 0.83 | 0.97 | 1.00 | 1.00 |
| 0.6 | 0.89 | 1.00 | 1.00 | 1.00 |

Correlated responses showed more dramatic changes when trait 2 was subject to subsampling, especially at the lowest intensity (Figure 3). In general, the expected relative response was boosted for trait 1, with some values well over 1, while the

expected relative response for trait 2 was depressed to less than 0.7. Thus, even when sometimes one of the correlated responses was superior to the value expected using the true genetic parameters the total direct gain was inferior. In most of the cases correlated response $\Delta_c G_1$ was superior to $\Delta_c G_2$. As with direct response, correlated response was dependent on heritability (especially of trait 2, Figure 3) and the effect of the genetic correlation was relevant only at low subsampling intensities.

**Figure 3:** Correlated relative responses to selection, expressed as the ratio of response to the predicted correlated response, using the true parameters and relative economic weights 2 (trait 1) and 1 (trait 2). Correlated responses for trait 1 (···•···) and 2 (···o···) for $h_1^2 = 0.1$, $h_2^2 = 0.4$ and $r_g = -0.3$. Correlated responses for trait 1 (–•–) and 2 (–o–) for $h_1^2 = 0.1$, $h_2^2 = 0.8$ and $r_g = -0.3$.




The effect of different relative economic weights was more important for lower subsampling intensities. Thus, the relative gain subsampling 3 trees was superior for the index with weights 2:1 than those with 1:1 and 1:2, reflecting the influence of better genetic parameter estimates for trait 1 (Table 4). However, when increasing subsampling intensity the differences tend to fade and finally disappear when sampling 30 trees. This is especially marked for high values of $h_2^2$ (0.8), where using only 9 trees greatly reduces the difference (Table 4).

**Table 4:** Average relative gain for different combinations of heritabilities and economic weights, considering 3, 9, 15 and 30 trees sampled for trait 2.

| Heritabilities | | Economic weights | Subsampling intensity (number of trees) | | | |
|---|---|---|---|---|---|---|
| $h_1^2$ | $h_2^2$ | $w_i$ | 3 | 9 | 15 | 30 |
| 0.1 | 0.4 | 2:1 | 0.76 | 0.87 | 0.93 | 0.99 |
| | | 1:1 | 0.68 | 0.84 | 0.91 | 1.00 |
| | | 1:2 | 0.66 | 0.84 | 0.92 | 0.99 |
| | 0.8 | 2:1 | 0.76 | 0.91 | 0.95 | 1.00 |
| | | 1:1 | 0.74 | 0.91 | 0.96 | 1.00 |
| | | 1:2 | 0.74 | 0.91 | 0.96 | 1.00 |
| 0.3 | 0.4 | 2:1 | 0.94 | 0.97 | 0.98 | 1.00 |
| | | 1:1 | 0.78 | 0.88 | 0.93 | 1.00 |
| | | 1:2 | 0.71 | 0.86 | 0.93 | 1.00 |
| | 0.8 | 2:1 | 0.88 | 0.94 | 0.96 | 1.00 |
| | | 1:1 | 0.78 | 0.91 | 0.95 | 1.00 |
| | | 1:2 | 0.77 | 0.92 | 0.96 | 1.00 |

***Truncation Subsampling***

Because of cost-saving concerns, sequential culling for different traits within one generation is a common practice in tree breeding. Firstly, all families are assessed and ranked for one trait and then a given percentage of the top families for that trait are sampled for the second one. This implies the use of non-random selection of the families to be assessed for the second trait, truncating the distributions. This situation was simulated for $h_1^2 = 0.1$, $h_2^2 = 0.8$ and $r_g = -0.3$, considering relative economic weights 2:1. The tests were generated as for random subsampling, the families ranked considering only trait 1, and then 15 individuals from the top 40% of the families were assessed for trait 2. Finally, the same analyses used for random subsampling were applied to the data sets.

Truncation selection of families for trait 1 led to larger skewness and bias for $\hat{h}_2^2$ and $\hat{r}_g$ (Table 1). An extreme case is that of $\hat{r}_g$, where the estimated parameter averaged 0.34 rather than -0.3. When predicting transmitting abilities the correlation between predicted and 'true' values was depressed to 0.62 (trait 1) and 0.37 (trait 2). These correlations were even inferior to those subsampling only 3 trees (Table 2), even though the total number of assessed trees was higher (1200 versus 600) with the consequent extra cost. The effect on relative direct gain was an overestimate, by a factor of 1.36, compared with the use of the 'true' parameters. This problem was caused mainly because of the large bias of the genetic correlation estimates. On the whole, even when the percentage of sampled families and individuals was generous for tree breeding standards, truncation subsampling was by far the worst scheme simulated for this study.

## Final Remarks

In general, increasing the number of individuals included in the analysis resulted in better parameter estimates and larger amounts of genetic gain. The lowest subsampling intensity proved to be generally inadequate for producing reliable estimates of genetic parameters, prediction of transmitting abilities and predictions of response to selection, all of which are important decision making tools for a breeding program. Raising the number of trees subsampled was subject to the Law of Diminishing Returns, with little effect over 15 trees. A potential for improved cost-efficiency is thus confirmed.

Concerning analytical tools, REML is becoming the preferred analysis method in forest genetics, mainly because of its statistical properties. HUBER et al. (1994) already showed its superior performance for various forest genetic experiments. The use of this procedure in the estimation of genetic parameters avoided completely the existence of out-of-bounds estimates, coping successfully with highly unbalanced data.

Even when the behaviour of random subsampling for trait 2 was fairly consistent compared with the use of 100% of the data (especially using 15 trees or more information), there is still room for improving the efficiency of the process. CAMERON and THOMPSON (1986) proposed 'elliptical selection' as an alternative method for characterising genetic and environmental contributions to the parent-offspring

relationships. This method concentrates the subsampling on the extremes of the distribution. It is a proposal worth considering for collateral relatives, especially if the main interest is the estimation of genetic parameters, and not ranking the parents. Nevertheless, given the multipurpose function of the progeny tests (WHITE 1987) a compromise may be needed between the optimal number of individuals for estimating genetic parameters and that for ranking the families.

The chosen subsampling intensity will depend on the expected additional profit derived from increased genetic gain. Optimisation of subsampling will depend on cost of assessment per individual, economic importance of the traits under study, expected gain of the subsampling scheme, scale of deployment of the selected material, and time frame between selection and harvesting the benefits on the plantations. Additionally, there could be different optima for parameter estimation, prediction of transmitting abilities and response to selection.

## Acknowledgements

The technical assistance of D.L. Johnson with AIREML is gratefully acknowledged. The help of R.J. Spelman in programming the simulations is recognised. Thanks to S. Kumar and C.T. Sorensson for valuable suggestions. L.A. Apiolaza was supported with NZODA and NZFRI scholarships during the completion of this study.

## Literature cited

BORRALHO, N.M.G., COTTERILL, P.P. and KANOWSKI, P.J. 1993. Breeding objectives for pulp production of *Eucalyptus globulus* under different industrial cost structures. *Canadian Journal of Forest Research* **23**: 648-656.

BROWN, G.H. 1969. An empirical study of the distribution of the sample genetic correlation coefficient. *Biometrics* **25**: 63-72.

BURDON, R.D. 1992. Genetic survey of *Pinus radiata*. 9: General discussion and implications for genetic management. *New Zealand Journal of Forestry Science* **22**: 274-298.

BURDON, R.D. and APIOLAZA, L.A. 1998. Short note: more generalised estimation of between trait genetic correlations using data from collateral relatives. *Silvae Genetica* **47**: 174-175.

CAMERON, N.D. and THOMPSON, R. 1986. Design of multivariate selection experiments to estimate genetic parameters. *Theoretical and Applied Genetics* **72**: 466-476.

CORNELIUS, J. 1994. Heritabilities and additive genetic coefficients of variation in forest trees. *Canadian Journal of Forest Research* **24**: 372-379.

GREAVES, B.L. and BORRALHO, N.M.G. 1996. The influence of basic density and pulp yield on the cost of eucalypt kraft pulping: a theoretical model for tree breeding. *Appita Journal* **49**: 423-426.

GREAVES, B.L., BORRALHO, N.M.G. and RAYMOND, C.A. 1997. Breeding objective for plantation eucalypts grown for production of kraft pulp. *Forest Science* **43**: 465-475.

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

HUBER, D.A., WHITE, T.L. and HODGE, G.R. 1994. Variance component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. *Theoretical and Applied Genetics* **88**: 236-242.

JOHNSON, D.L. and THOMPSON, R. 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse techniques and average information. *Journal of Dairy Science* **78**: 449-456.

JOHNSON, M. E. 1987. Multivariate Statistical Simulation. Wiley & Sons, New York.

LIU, B.-H., KNAPP, S.J. and BIRKES, D. 1997. Sampling distributions, biases, variances, and confidence intervals for genetic correlations. *Theoretical and Applied Genetics* **94**: 8-19.

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

ROBERTSON, A. 1959. The sampling variance of the genetic correlation coefficient. *Biometrics* **15**: 469-485.

ROFF, D.A. and PREZIOSI, R. 1994. The estimation of the genetic correlation: the use of the jackknife. *Heredity* **73**: 544-548.

SHELBOURNE, C.J.A., APIOLAZA, L.A., JAYAWICKRAMA, K.J.S. and SORENSSON, C.T. 1997. Developing breeding objectives for radiata pine in New Zealand. P 160-168 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

SNEDECOR, G.W. and COCHRAN, W.G. 1980. Statistical methods, 7th edition. The Iowa State University Press, Iowa.

THOMPSON, R. and MEYER, K. 1986. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livestock Production Science* **15**: 299-313.

VAN VLECK, L.D. 1993. Selection index and introduction to mixed model methods. CRC Press, Florida.

VAN VLECK, L.D. 1994. Algorithms for simulation of animal models with multiple traits and with maternal and non-additive genetic effects. *Brazilian Journal of Genetics* **17**: 53-57.

VAN VLECK, L.D. and HENDERSON, C.R. 1961. Empirical sampling estimates of genetic correlations. *Biometrics* **17**: 359-371.

WHITE, T.L. 1987. A conceptual framework for tree improvement programs. *New Forests* **1**: 325-342.

## Appendix

The bivariate prediction of transmitting abilities of a parent for trait i (in this case i = 1,2) has the form:

$$\hat{f}_i = \hat{c}_{i1}\overline{P}_1 + \hat{c}_{i2}\overline{P}_2$$

where $\overline{P}_i$ is the average phenotypic information for trait i, expressed as deviation from the generalised least square estimation of the overall mean of that trait. The values of $\hat{\mathbf{c}}_i$ are:

$$\hat{\mathbf{c}}_i = \mathbf{P}^{-1}\,\mathbf{q}_i$$

where

$$\mathbf{P} = \begin{bmatrix} V(\overline{P}_1) & Cov(\overline{P}_1,\overline{P}_2) \\ Cov(\overline{P}_1,\overline{P}_2) & V(\overline{P}_2) \end{bmatrix} \text{ and } \mathbf{q}_i = \begin{bmatrix} Cov(\overline{P}_1, f_i) \\ Cov(\overline{P}_2, f_i) \end{bmatrix}$$

moreover

$$V(\overline{P}_1) = \frac{\sigma_{P_1}^2}{n_1}\left[1 + (n_1 - 1)¼\, h_1^2\right],\quad V(\overline{P}_2) = \frac{\sigma_{P_2}^2}{n_2}\left[1 + (n_2 - 1)¼ h_2^2\right],\quad Cov(\overline{P}_1, f_i) = \frac{\sigma_{a_1 a_i}}{2},$$

$$Cov(\overline{P}_2, f_i) = \frac{\sigma_{a_2 a_i}}{2}, \text{and } Cov(\overline{P}_1, \overline{P}_2) = \frac{\sigma_{P_1 P_2}}{n_1} + \frac{(n_2 - 1)¼\sigma_{a_1 a_2}}{n_1}$$

where $n_i$ is the number of observations per family for trait i, $\sigma_{P_i}^2$ is the phenotypic variance of trait i, $\sigma_{P_i P_j}$ is the phenotypic covariance between traits i and j, and $\sigma_{a_i a_j}$ is the additive genetic covariance between traits traits i and j.

The bivariate predicted transmitting abilities for traits 1 and 2 ($\hat{f}_1$ and $\hat{f}_2$) are combined into an index I. This index is used as a prediction of the breeding objective H, which includes the transmitting abilities ($f_1$ and $f_2$) of the same traits as those in the index. Thus:

$H = w_1 f_1 + w_2 f_2$

$I = w_1 \hat{f}_1 + w_2 \hat{f}_2$

where $w_i$ are the relative economic weights of the traits.

Expected correlated response of trait i to selection on index I is (VAN VLECK 1993):

$\Delta_c G_i = i \, Cov(I,f_i) \, V(I)^{-1/2}$

where $Cov(I,f_i)$ is the covariance between the index and the transmitting abilities for trait i, and V(I) is the variance of the selection index.

Considering **S** the matrix of predicted transmitting abilities variances and covariances, $\mathbf{T}_i$ the vector of covariances between predicted transmitting abilities for both traits and the true genetic values for trait i, and a selection intensity *i*:

$$\mathbf{S} = \begin{bmatrix} \hat{\mathbf{c}}_1'\mathbf{P}\hat{\mathbf{c}}_1 & \hat{\mathbf{c}}_1'\mathbf{P}\hat{\mathbf{c}}_2 \\ \hat{\mathbf{c}}_1'\mathbf{P}\hat{\mathbf{c}}_2 & \hat{\mathbf{c}}_2'\mathbf{P}\hat{\mathbf{c}}_2 \end{bmatrix} \text{ and } \mathbf{T}_i = \begin{bmatrix} \hat{\mathbf{c}}_1'\mathbf{q}_i \\ \hat{\mathbf{c}}_2'\mathbf{q}_i \end{bmatrix}$$

correlated response of trait i reduces to:

$\Delta_c G_i = i \, \mathbf{w}' \, \mathbf{T}_i \, (\mathbf{w}' \, \mathbf{S} \, \mathbf{w})^{-1/2}$

while direct response to selection is (VAN VLECK 1993):

$\Delta G_H = w_1 \Delta_c G_1 + w_2 \Delta_c G_2$

# CHAPTER FOUR

# ANALYSIS OF LONGITUDINAL DATA FROM PROGENY TESTS: SOME MULTIVARIATE APPROACHES

Luis A. Apiolaza[1,2] and Dorian J. Garrick[1]

[1]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand. [2]New Zealand Forest Research Institute, Private Bag 3020, Rotorua, New Zealand.

FOREST SCIENCE (IN PRESS)

## 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 changes of heritability and genetic correlations with age. This paper 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 genetic 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 relationship between 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 paper 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 weighted basic wood density (in $kg/m^3$) at 4 ages (5, 10, 15 and 20 years) from Radiata Pine increment cores.

## Introduction

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 in this paper, 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).

Longitudinal data allow modeling changes of heritability and genetic correlations with age. 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, its 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 are 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 paper 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 relationship between 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 ($e_i$):

$$y_i = \mathbf{x}_i' \mathbf{b} + a_i + e_i \quad [1]$$

where **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 1s and 0s 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:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \quad [2]$$

where **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 **X** (obtained by stacking $\mathbf{x}_i'$ for all trees) and **Z** links observations to **b** and **a**, respectively. The vector of expected values and the dispersion matrices are defined by:

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

$$\mathrm{Var}[\mathbf{a}] = \mathbf{G} = \mathbf{A}_N \sigma_a^2,\ \mathrm{Var}[\mathbf{e}] = \mathbf{R} = \mathbf{I}\sigma_e^2 \text{ and } \mathrm{Var}[\mathbf{y}] = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \quad [3]$$

where $\mathbf{A}_N$ is the numerator relationship matrix, which describes the additive genetic relationship among individuals with records (see MRODE 1996 chapter 2 for a detailed explanation). In addition, **I** is an identity matrix, $\sigma_a^2$ is the additive genetic variance and $\sigma_e^2$ is the error variance. Random effects **a** and **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 **y**, **a** and **e** follow a multivariate normal distribution, and provided **G** and **R** are positive definite, best linear unbiased predictions (BLUP, HENDERSON 1984) of the breeding values of the individuals can be calculated using Henderson's mixed model equations (HENDERSON 1984):

$$\begin{bmatrix} \mathbf{X`R^{-1}X} & \mathbf{X`R^{-1}Z} \\ \mathbf{Z`R^{-1}X} & \mathbf{Z`R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X`R^{-1}y} \\ \mathbf{Z`R^{-1}y} \end{bmatrix} \quad [4]$$

where **G** and **R** are functions of $\sigma_a^2$ and $\sigma_e^2$ 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 $\sigma_a^2$ and $\sigma_e^2$, subject to the constraints that these parameters are within the parameter space (i.e., non negative and sum to the total phenotypic variance):

$$\text{Log L} = -\tfrac{1}{2} [\text{con} + \log |\mathbf{G}| + \log |\mathbf{R}| + \log |\mathbf{C}| + \mathbf{y'\,P\,y}] \quad [5]$$

where con is a constant **G** and **R** are as from equation 3, **C** is the coefficient matrix of equation 4, **P** is the projection matrix $\mathbf{V^{-1}} - \mathbf{V^{-1}\,X\,(X'\,V^{-1}\,X)^{-}\,X'\,V^{-1}}$, and $\mathbf{(X'\,V^{-1}\,X)^{-}}$ represents a generalized inverse of $\mathbf{(X'\,V^{-1}\,X)}$. The matrix **P** absorbs the fixed effects and accounts for information about **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 $\mathbf{b} = [\mathbf{b'}_{\text{trait1}}\ \mathbf{b'}_{\text{trait2}}\ \ldots\ \mathbf{b'}_{\text{traitm}}]'$ 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 non-inbred individual are defined by:

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

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

In the multivariate approach $\mathbf{G}_0$ and $\mathbf{R}_0$ represent the m x 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 $\sigma_{e_{jk}}$. Random effects $\mathbf{a}_i$ and $\mathbf{e}_i$ are assumed uncorrelated. This model can be easily expanded to include additional 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 $\mathbf{e} = [\mathbf{e}_1'\ \mathbf{e}_2'\ \ldots\ \mathbf{e}_N']'$. Additionally, $\mathbf{X} = [\mathbf{X}_1'\ \mathbf{X}_2'\ \ldots\ \mathbf{X}_N']'$ and $\mathbf{Z} = \Sigma_{\oplus}\mathbf{Z}_i$, where $\Sigma_{\oplus}$ represents direct sum operation. Consequently, $\mathbf{G} = \mathbf{A}_N \otimes \mathbf{G}_0$ and $\mathbf{R} = \Sigma_{\oplus}\mathbf{R}_i$, where $\otimes$ denotes direct product (see Appendix 1 and SEARLE 1982 (chapter 10) for a detailed description of $\Sigma_{\oplus}$ and $\otimes$ operations) and $\mathbf{R}_i$ is the residual covariance matrix for individual i. Hence, the expected value and dispersion matrices are:

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

$$Var[\mathbf{a}] = \mathbf{G} = \mathbf{A}_N \otimes \mathbf{G}_0,\ Var[\mathbf{e}] = \mathbf{R} = \Sigma_{\oplus}\mathbf{R}_i \text{ and } Var[\mathbf{y}] = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R} \quad [8]$$

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 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 4 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 usual 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 ($\alpha_j$), which are considered independent of the past measurement:

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

where $\rho_{jk}$ 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}→g, {b, e}→h, and {c}→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).

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

The relationship between successive breeding values is similar to equation 12, but $\rho$ is the same for all observations separated by a lag k:

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

This assumption may not be applicable across different growth stages, where development in one year of, say, early growth can be very different from that of one year 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 $\mathbf{M} = \mathbf{S}\,\mathbf{C}_{AR}\,\mathbf{S}$, further reducing to one the number of covariances to estimate.

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

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 ($\alpha_j$):

$$a_{ij} = \rho\, a_{i\,j-1} + \alpha_j \quad [16]$$

If the correlation ($\rho$) is a function of a unique value and the lag between the measurements, the relationship between successive breeding values for individual i is:

$$\begin{aligned} a_{i\,j-1} &= \rho\, a_{i\,j-2} + \alpha_{j-1} \\ a_{i\,j-2} &= \rho\, a_{i\,j-3} + \alpha_{j-2} \\ &\vdots \\ a_{i\,j-k+1} &= \rho\, a_{i\,j-k} + \alpha_{j-k+1} \end{aligned} \quad [17]$$

and substituting every breeding value of equation 17 in the preceding one we obtain:

$$a_{ij} = \rho^{|j-k|}\, a_{i\,j-k} + \alpha'_j \quad [18]$$

where $\alpha'_j$ 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 $\rho = e^{-k\ lag}$ (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 $\mathbf{C}_{RE} = \mathbf{J}$.

$$\mathbf{C}_{RE} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad [19]$$

As all rows are identical $\mathbf{G}_0$ is singular limiting the use of mixed model equations (equation 4) in their normal form. A solution for this problem is the regularly used alternative 'univariate' representation of the model:

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

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 similar to the univariate case, while phenotypic variance now includes permanent environment variance:

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

$$Var[\mathbf{h}] = \mathbf{H} = \mathbf{I}\,\sigma_h^2 \text{ and } Var[\mathbf{y}] = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{W}\,\mathbf{H}\,\mathbf{W}' + \mathbf{R} \quad [21]$$

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 for 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 are no genetic or residual associations 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.

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

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:

$$\mathbf{y}_i = f_b(\mathbf{t}) + f_{a_i}(\mathbf{t}) + f_{e_i}(\mathbf{t}) + \varepsilon_i \quad [23]$$

where $f_b(\mathbf{t})$, $f_{a_i}(\mathbf{t})$ and $f_{e_i}(\mathbf{t})$ represent possibly different functions modeling fixed effects, additive genetic effects and residuals respectively; and $\varepsilon_i$ 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). The emphasis in this paper is on modeling the additive genetic covariance matrix (**G**), with random regressions used for $\mathbf{a}_i$ while other terms are considered unstructured and the subindex for $f(t)$ is dropped. If $\mathbf{a}_i = f(\mathbf{t})$ with **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 ($a_{ij}$):

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

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

$$\mathbf{a}_i = f(\mathbf{t}) = \mathbf{Q}_i \lambda_i \quad [25]$$

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

$$Q_i = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{n1} \\ z_{02} & z_{12} & \cdots & z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0m} & z_{1m} & \cdots & z_{nm} \end{bmatrix} \quad [26]$$

Therefore equation 6 can be represented as:

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{b} + \mathbf{Q}_i \lambda_i + \mathbf{e}_i \quad [27]$$

$$\text{with } \mathrm{Var}[\mathbf{Q}_i \lambda_i] = \mathbf{Q}_i \Lambda_0 \mathbf{Q}_i' \quad [28]$$

where $\Lambda_0$ is the covariance matrix of the random coefficients ($\lambda_i$). Because different regression coefficients are calculated for every individual (and these coefficients are considered as random effects), this model is called a random regression 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 page 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), 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.

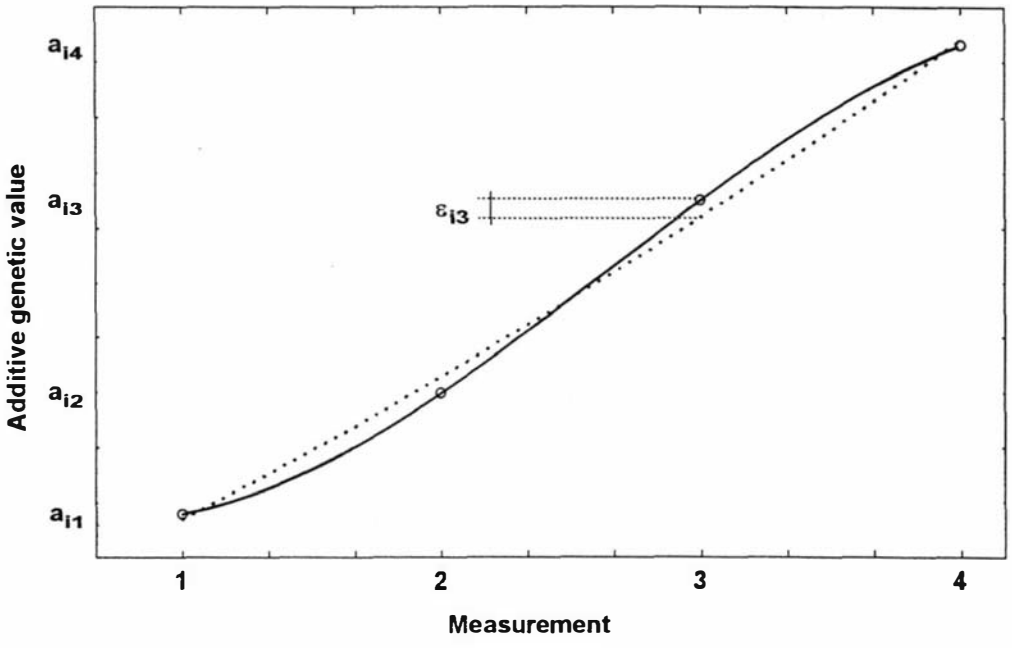
## Relationship between unstructured and random regression models

Two linear models $m_1$ and $m_2$ are considered equivalent when their expected values and variances are identical (HENDERSON 1984 page 6):

$$E[m_1] = E[m_2]$$
$$Var[m_1] = Var[m_2] \quad [29]$$

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 4 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 $\varepsilon_{i3}$.




Using a US multivariate approach, i.e. model equation 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 $\lambda_i$ 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 page 276), the product $\mathbf{Q}_i\,\lambda_i$ on equation 31 is:

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

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,

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

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

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:

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

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

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

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 ($\varepsilon_i$) 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 $\varepsilon_i$.

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 a 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 dependant 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 $\Phi$ is a matrix of orthogonal polynomials (Legendre polynomials in Kirkpatrick's work) with columns $\phi$, $\mathbf{G}_0$ is a covariance matrix (e.g. additive genetic), and $\mathbf{U}_0$ is the covariance matrix of the polynomial coefficients then:

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

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

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 $\Phi$

(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. While 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:

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

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:

$$AIC = -2\ LogL + 2\ p \quad [40]$$

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(**X**) in p, because it will not affect the differences in 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 non-nested 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 year 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:

$$wbd_j = \frac{\sum_{i=1}^{j} bd_i\ \Delta_i}{\sum_{i=1}^{j} \Delta_i} \quad (41)$$

where $bd_i$ is the average basic density of ring i and $\Delta_i$ 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 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 were 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.

**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 | -4886.90 | 9813.80 |
| BC | 7 + 10 | -4891.03 | 9816.06 |
| AR | 5 + 10 | -4892.71 | 9815.42 |
| RE | 1 + 2 | -6327.04 | 12660.08 |
| UC | 4 + 4 | -6715.32 | 13446.64 |
| RRr | 6 + 10 | -4888.27 | 9808.54 |

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 $\hat{\mathbf{R}}_0$:

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

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

**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 ($\sigma_p^2$) and residual correlations ($r_e$, below diagonal).

| Age (years) | $h^2$ | $\sigma_p^2$ | Age (years) | | |
|---|---|---|---|---|---|
| | | | 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 | 1052.468 | | | |
| 10 | 0.567 | 1052.468 | 0.180 | | |
| 15 | 0.567 | 1052.468 | 0.180 | 0.180 | |
| 20 | 0.567 | 1052.468 | 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.

**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.

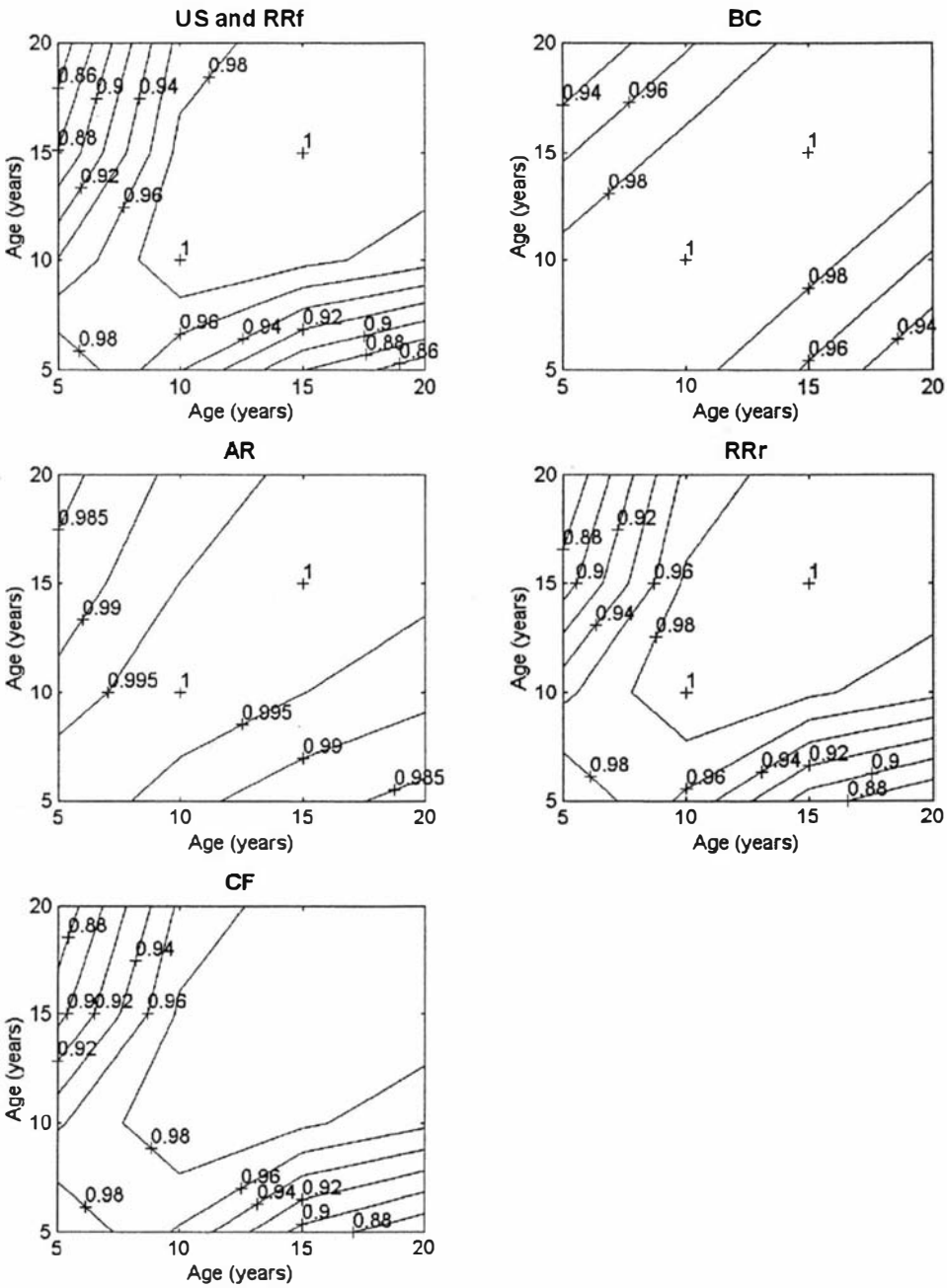


Table 2 presents genetic parameters estimates from the different models. As expected the US and RRf (fitting a third-order orthogonal polynomial for each tree) models produce identical estimates of genetic parameters. The RRr model, which fits a second-order orthogonal polynomial for each tree, has a very similar fit with only 6 parameters in the $\mathbf{G}_0$ matrix.

In general heritability estimates do not differ substantially between the models; however, the estimates for age 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).

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}→0.988, {0.881, 0.968}→0.958 and {0.846}→0.917 from the $C_{US}$ (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.

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

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 cannot 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 20) 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., 6 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 Cov(x,y) = [Cov(x+y) – Var(x) – 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 5 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 4 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 paper 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.

## Acknowledgements

Luis Apiolaza was funded by NZODA and NZFRI scholarships. The dataset used in the example was compiled by Paul Jefferson, based in results from a densitometry analysis. The NZ Radiata Pine Breeding Co-operative 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.

## Literature cited

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions of Automatic Control* **19**: 716-723.

APIOLAZA, L.A., GILMOUR, A.R., and GARRICK, D.J. 1999. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. *Canadian Journal of Forest Research* **30**: 645-654.

BORRALHO, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. P. 141-145 *in* POTTS, B.M., BORRALHO, N.M.G., REID, J.B., CROMER, R.N., TIBBITS, W.N., and RAYMOND, C.A. "Eucalypts plantations: improving fibre yield and quality". Proceedings of CRC-IUFRO Conference, 19-24 February, Hobart, Tasmania, Australia.

BURDON, R.D. 1989. Early selection in tree breeding: principles for applying index selection and inferring input parameters. *Canadian Journal of Forest Research* **19**: 499-504.

CNAAN, A., LAIRD, N.M., and SLASOR, P. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* **16**: 2349-2380.

COELLI, K.A., GILMOUR, A.R., and ATKINS, K.D. 1998. Comparison of genetic covariance models for annual measurements of fleece weight and fibre diameter. P. 31-34, Vol. 24 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production. 11-16 January, Armidale, NSW, Australia.

DIETERS, M.J., WHITE, T.L., LITELL, R.C. and HODGE, G.R. 1994. Application of approximate variances of variance components and their ratios in genetic tests. *Theoretical and Applied Genetics* **91**: 15-24.

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

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

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. *American Naturalist* **86**: 293-298.

GABRIEL, K.R. 1962. Ante-dependence analysis of an ordered set of variables. *Annals of Mathematical Statistics* **33**: 201-212.

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

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

GREGOIRE, T.G., SCHABENBERGER, O., and BARRET, J.P. 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. *Canadian Journal of Forest Research* **25**: 137-156.

HAND, D., and CROWDER, M. 1996. Practical longitudinal data analysis. Chapman and Hall, London, United Kingdom.

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

HUBER, D.A., WHITE, T.L., and HODGE, G.R. 1994. Variance component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. *Theoretical and Applied Genetics* **88**: 236-242.

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

JENNRICH, R.I., and SCHLUCHTER, M.D. 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.

KIRKPATRICK, M., and HECKMAN, N. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**: 429-450.

KIRKPATRICK, M., LOFSVOLD, D. and BULMER, M. 1990. Analysis of inheritance, selection and evolution of growth trajectories. *Genetics* **124**: 979-993.

KIRKPATRICK, M., HILL, W.G., and THOMPSON, R. 1994. Estimating the covariance structures for traits during growth and ageing, illustrated with lactation in dairy cattle. *Genetical Research* **64**: 59-67.

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

LINDSEY, J.K., and JONES, B. 1998. Choosing among generalized linear models applied to medical data. *Statistics in Medicine* **17**: 59-68.

LOUIS, T.A. 1988. General methods for analysing repeated measures. *Statistics in Medicine* 7: 29-45.

MAGNUSSEN, S., and KREMER, A. 1993. Selection for an optimum tree growth curve. *Silvae Genetica* **42**: 322-335.

MEYER, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. *Genetics, Selection, Evolution* **30**: 221-240.

MEYER, K., and HILL, W.G. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science* **47**: 185-200.

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

NETER, J., and WASSERMAN, W. 1974. Applied linear statistical models. Richard D. Irving, Homewood.

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

QUAAS, R.L., ANDERSON, R.D., and GILMOUR, A.R. 1984. BLUP School Handbook. Animal Genetics and Breeding Unit, University of New England, Australia.

SEARLE, S.R. 1982. Matrix algebra useful for statistics. John Wiley and Sons, New York.

SHAW, R.G. 1991. The comparison of quantitative genetic parameters between populations. *Evolution* **45**: 143-151.

SHELBOURNE, C.J.A., and LOW, C.B. 1980. Multi-trait index selection and associated genetic gains of Pinus radiata progenies at five sites. *New Zealand Journal of Forest Science* **10**: 307-324.

SORIA, F., BASURCO, F., TOVAL, G., SILIÓ, L., RODRÍGUEZ, M.C., and TORO, M.A. 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 conference on Silviculture and Improvement of Eucalypts. Salvador, Brazil.

VAN DER WERF, J.H.J., and SCHAEFFER, L. 1997. Random regression in animal breeding. Course notes. CGIL, 25-28 Jun. 1997, Guelph, Canada.

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

WADA, Y., and KASHIWAGI, N. 1990. Selecting statistical models with information statistics. *Journal of Dairy Science* **73**: 3575-3582.

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

WHITE, I.M.S., THOMPSON, R., and BROTHERSTONE, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science* **88**: 632-638.

## Appendix: direct sum and direct product

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

$$\Sigma_{\oplus}\mathbf{A}_i = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_n \end{bmatrix} = \text{diag}\{\mathbf{A}_i\} \quad [A1]$$

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**:

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

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

Example:

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

# CHAPTER FIVE

# VARIANCE MODELLING OF LONGITUDINAL HEIGHT DATA FROM A *PINUS RADIATA* PROGENY TEST

Luis A. Apiolaza[1,2], Arthur R. Gilmour[3], Dorian J. Garrick[1]

[1]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand. [2]New Zealand Forest Research Institute, Private Bag 3020, Rotorua, New Zealand. [3]New South Wales Agriculture, Orange 2800, Australia.

CANADIAN JOURNAL OF FOREST RESEARCH **30**: 645-654.

## Abstract

Variance components were estimated using alternative structures for the additive genetic covariance matrix ($\mathbf{G}_0$), for height (m) of trees measured at 10 unequally spaced ages in an open-pollinated progeny test. These structures reflected unstructured, autoregressive, banded correlation and random regressions models. The residual matrix ($\mathbf{R}_0$) was unstructured, and the block and plot strata matrices were autoregressive. The best model for $\mathbf{G}_0$ considering the likelihood value and number of parameters was the autoregressive correlation form with age specific variances and time on a natural logarithm basis. The genetic correlations between successive measures ranged from 0.93 at age 1 to 0.99 at age 14 years. Heritability increased with age from 0.09 (age 1) to 0.24 (age 7) and then declined to 0.13 at age 15. Heritabilities from the unstructured model were similar, while heritabilities assuming banded correlations were lower after age 7. The covariance structure implicit in the random regressions model was considered unsatisfactory. Using structures in $\mathbf{G}_0$ facilitated model fitting and convergence of the likelihood maximisation algorithm. Fitting a structured matrix that reflects the relationships present in repeated measures may overcome problems of non-positive definiteness of unstructured matrices from longitudinal data, especially when heritability is small.

## Introduction

The success of tree breeding programmes relies on their ability to identify and deploy superior trees, progressively increasing profit. Decisions on how, when and what to select are made (or should be made) taking into account genetic and economic information. Selection in tree breeding programmes is based upon genetic information generated from progeny tests.

The net present value of genetic gain depends on the total genetic gain ($\Delta G$) and the time when improved material is deployed and costs (L) are incurred (NEWMAN and WILLIAMS 1991). Both $\Delta G$ and L contain constraints and opportunities for breeding programmes to make rapid gains. Faster gains can be achieved by increasing the selection differentials and/or the accuracy of prediction. Furthermore, using overlapping generations and early selection can reduce long generation intervals. Efficient selection at an early age requires

high correlation with rotation-age production traits and reasonably high heritabilities of both. Knowledge of the expected covariance structure across ages enables prediction of the response to early selection.

'Longitudinal' data arise when individuals are assessed for the same outcome at several ages (DIGGLE et al. 1994, see CNAAN et al. 1997 for a review). Breeders use longitudinal data from progeny tests to compare development patterns of genotypes and look at changes in genetic parameters over time. Understanding the genetics of development allows determination of the optimum evaluation time(s) for fine-tuning breeding programmes and the use of multiple assessments for genetic evaluations. One of the features of longitudinal data is the covariance that exists between observations of the same individual (DIGGLE et al. 1994, HAND and CROWDER 1996). Covariance matrices typically contain pattern or structure, which can be modelled with a reduced number of parameters. DIGGLE et al. (1994) identify three sources of variation in longitudinal data: serial correlations, random effects and measurement error. These act simultaneously.

Previous researchers in tree breeding have used several variance models for longitudinal data. Early studies used univariate analysis at each age, sometimes fitting a curve through the estimates to smooth and interpolate the results (e.g. FOSTER 1986). This procedure may produce unbiased estimates of heritability in absence of selection but ignores the dependence (covariance) between times. Other studies used bivariate analyses of each pair of times (e.g. BALOCCHI et al. 1993), increasing the understanding of the association between measures. Another procedure uses the correlations between family means of the same genetic material grown at different sites, i.e. based on Type B correlation (BURDON 1977, see HODGE and WHITE 1992 for an extensive application), with the intention of avoiding error correlation between the measures.

Longitudinal analyses are more efficient using all available information, especially when missing observations are a problem. Progeny tests, as other long-term forestry experiments, do not maintain the original design. Even when a test starts as a balanced experiment, mortality generates temporal imbalance; hence early measures contain more records than later ones (GREGOIRE et al. 1995). Some authors choose to eliminate

temporal unbalance, keeping only individuals with a full history of measures (e.g. BALOCCHI et al. 1993), but this approach omits useful data and does not consider that biases may arise if mortality is not random.

There are recent attempts with forest trees to use two contrasting models for genetic correlation: a repeatability (univariate) model (e.g. WEI and BORRALHO 1996) and a full multivariate model (e.g. WEI and BORRALHO 1998). The repeatability model assumes that all measures represent the same trait. This implies a genetic correlation of one between all pairs of records, equal variance for all records and equal environmental correlation between all pairs of records. The model can be represented by $\mathbf{G} = \sigma_a^2\mathbf{J}$ and $\mathbf{R} = \sigma_e^2(\mathbf{I} + r\mathbf{J})$ where **I** is the identity matrix, **J** a square matrix with all elements equal to 1, $\sigma_a^2$ is the additive genetic variance component, $\sigma_e^2$ is the error variance component and r/(1+r) is the correlation between residuals. With height increasing with age over many years, the equal variance and genetic correlation assumptions are often unrealistic, although variance heterogeneity may be crudely removed through standardisation. In contrast, the full multivariate model considers each age as the realisation of a different trait. It was originally applied to balanced and complete data, but modern computing techniques allow its application to incomplete data sets.

Unstructured covariance matrices in a full multivariate analysis are feasible and reasonable with a small number t of successive measures. However, since these matrices have t(t-1)/2 covariance components, more measures implies a large number of poorly estimated parameters and may be considered an overparameterisation (HAND and CROWDER 1996). That is, there may not be enough information in the data to estimate each variance and covariance with sufficient accuracy for the resulting matrix to be coherent (positive definite), and the likelihood maximisation algorithm may even fail to converge. This is less a problem with some structured matrices, making appealing the assumption of more parsimonious covariance structures as t increases.

Some researchers have recognised the need for structured covariance matrices, yet only in isolated cases. For example QUAAS (1984, page 34) proposed the use of an autoregressive error structure in a repeatability model, relaxing the equal correlation

assumption. KREMER (1992) explicitly recognised the role of error serial correlation for the analysis of height increments. COELLI et al. (1998) compared several different structures for the multivariate analysis of additive genetic effects of repeated measures in a sheep breeding context.

An alternative approach for modelling covariance structures, regression models with random coefficients, was introduced by RAO (1965) in the context of growth models. LAIRD and WARE (1982) generalised the theory to include mixed models, with fixed parameters at the population level and random parameters at the individual level. Under this framework, unbalanced and incomplete data sets are readily handled, and the correlation among successive measures is implicitly modelled by the random regressions (LOUIS 1988, VONESH and CHINCHILLI 1997). SCHAEFFER and DEKKERS (1994), JAMROZIK and SCHAEFFER (1997) and JAMROZIK et al. (1997) applied random regression models to the analysis of lactation records; while GREGOIRE et al. (1995) advocated the use of random regressions to model growth in permanent plots in forest mensuration. The use of random regressions in these studies allowed for individuals with heterogeneous ages to be included in the analyses, and a reduction of computational requirements compared to unstructured multivariate analyses. Random regression models directly define covariance functions that are the continuous (infinitesimal) equivalent of a covariance matrix for a given trait and fixed ages (KIRKPATRICK et al. 1990, KIRKPATRICK et al. 1994). These functions permit us to calculate the covariance between any two ages, as can also be done with the distance based autocorrelation model. So, the association among measurements may be modelled either through random regressions or through specification of covariance structures. In some cases both methods are used together (e.g. CHI and REINSEL 1989, JONES 1990).

In this paper we model longitudinal data from a progeny test. We compare estimated genetic parameters obtained from traditional approaches with those from various variance models under the general mixed model. First we introduce the general model in a tree breeding context. Then we present alternative structures for the additive genetic covariance matrix. Finally, we discuss opportunities and limitations for the use of these models.

## Materials and methods

### *Data set*

The forestry company Bosques Arauco S.A. established trial FA8102 in the 'VIII Región' of Chile in 1981 to progeny test 45 radiata open-pollinated first-generation selections with 9 controls. These were planted in 5 tree plots within 8 randomised complete blocks, a total of 2160 trees. Control plots were planted with mixed seed of unknown pedigree and have been omitted from the analysis. Trees that where suppressed by early competition never reached 5 cm of DBH, and were also omitted. Consequently, a total of 1526 trees in 353 plots were included in the analysis, each tree being from seed collected from one of 45 mother trees.

**Table 1:** Summary statistics by age. The percentage of trees refers to the individuals included in the analyses relative to the number initially established (1721, without considering controls and fillers) discounting natural mortality, mechanical damage and inconsistent measures.

| Age | Number of individuals | Percentage of trees (%) | Height mean (m) (standard deviation) |
|---|---|---|---|
| 1 | 1522 | 88.4 | 0.48 (0.12) |
| 2 | 1525 | 88.6 | 1.00 (0.22) |
| 4 | 1525 | 88.6 | 2.98 (0.60) |
| 5 | 1524 | 88.6 | 4.60 (0.85) |
| 6 | 1511 | 87.8 | 6.45 (1.02) |
| 7 | 1515 | 88.0 | 8.43 (1.11) |
| 8 | 1510 | 87.7 | 10.43 (1.22) |
| 9 | 1505 | 87.4 | 12.11 (1.36) |
| 12 | 1351 | 78.5 | 17.97 (1.79) |
| 15 | 1284 | 74.6 | 22.34 (2.46) |

The trees were assessed for height at 1, 2, 4, 5, 6, 7, 8, 9, 12 and 15 years of age. In the case of trees with mechanical damage (as those broken by wind at age 12), observations after the damage occurred were eliminated. Summary statistics by age are presented in Table 1. Small fluctuations of the number of trees between ages 1-2 and 6-7 are caused

by the elimination of inconsistent assessments. It is only after 9 years that there is appreciable mortality among the trees. The regression of log(standard deviation) on log(mean height) has a slope of 0.74±0.03 suggesting that a power transformation of the data to height$^{0.26}$ would stabilise the variance (BOX and COX, 1964). We will however analyse the data on the measurement scale using models with heterogeneous variances, to account for the increase of variance with age.

***General linear mixed model***

An individual tree ('animal') linear mixed-effects model equation for longitudinal data of tree i can be expressed as:

$$\mathbf{y}_i = \mathbf{X}_i\,\mathbf{m} + \mathbf{Z}_{Bi}\,\mathbf{b} + \mathbf{Z}_{Pi}\,\mathbf{p} + \mathbf{Z}_{Ti}\,\mathbf{a}_i + \mathbf{e}_i \qquad [1]$$

where $\mathbf{y}_i$ is the vector of $s_i$ observations for the individual indexed by age, $\mathbf{m}$ is the vector of fixed effects (which may include regression coefficients at population level), $\mathbf{b}$ is the vector of random block by age effects, $\mathbf{p}$ is the vector of random plot by age effects, $\mathbf{a}_i$ is the vector of individual random additive genetic effects, and $\mathbf{e}_i$ is the vector of random residuals. $\mathbf{X}_i$, $\mathbf{Z}_{Bi}$, $\mathbf{Z}_{Pi}$, and $\mathbf{Z}_{Ti}$ are incidence matrices relating $\mathbf{m}$, $\mathbf{b}$, $\mathbf{p}$ and $\mathbf{a}_i$ to $\mathbf{y}_i$. Thus the expected value and dispersion matrices assuming a multivariate normal distribution (MVN) are:

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{b} \\ \mathbf{p} \\ \mathbf{a}_i \\ \mathbf{e}_i \end{bmatrix} \sim \text{MVN}\left( \begin{bmatrix} \mathbf{X}_i\mathbf{m} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V}_i & \mathbf{Z}_{Bi}\mathbf{B}_0 & \mathbf{Z}_{Pi}\mathbf{P}_0 & \mathbf{Z}_{Ti}\mathbf{G}_0 & \mathbf{R}_0 \\ \mathbf{B}_0\mathbf{Z}_{Bi}' & \mathbf{B}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{P}_0\mathbf{Z}_{Pi}' & \mathbf{0} & \mathbf{P}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_0\mathbf{Z}_{Ti}' & \mathbf{0} & \mathbf{0} & \mathbf{G}_0 & \mathbf{0} \\ \mathbf{R}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_0 \end{bmatrix} \right) \qquad [2]$$

where $\mathbf{B}_0$, $\mathbf{P}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$ are the block, plot, additive genetic and residual covariance matrices respectively and $\mathbf{0}$ is a null matrix (with all elements equal to 0). The corresponding characteristic elements (for measures j and k) are $\sigma_{b_{jk}}$, $\sigma_{p_{jk}}$, $\sigma_{a_{jk}}$ and $\sigma_{e_{jk}}$. The number of observations per individual may vary in which case the corresponding rows and columns of $\mathbf{R}_0$ are deleted. Finally, the phenotypic covariance matrix is:

$$\mathbf{V}_i = \mathbf{Z}_{Bi}\,\mathbf{B}_0\,\mathbf{Z}_{Bi}' + \mathbf{Z}_{Pi}\,\mathbf{P}_0\,\mathbf{Z}_{Pi}' + \mathbf{Z}_{Ti}\,\mathbf{G}_0\,\mathbf{Z}_{Ti}' + \mathbf{R}_0 \qquad [3]$$

Extending the model equation to the n subjects of a progeny test we obtain:

$$\mathbf{y} = \mathbf{X}\,\mathbf{m} + \mathbf{Z}_B\,\mathbf{b} + \mathbf{Z}_P\,\mathbf{p} + \mathbf{Z}_T\,\mathbf{a} + \mathbf{e} \qquad [4]$$

for $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n')'$, $\mathbf{a} = (\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_n')'$, $\mathbf{e} = (\mathbf{e}_1', \mathbf{e}_2', \ldots, \mathbf{e}_n')'$, $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')'$, $\mathbf{Z}_B = (\mathbf{Z}_{B1}', \mathbf{Z}_{B2}', \ldots, \mathbf{Z}_{Bn}')'$, $\mathbf{Z}_P = (\mathbf{Z}_{P1}', \mathbf{Z}_{P2}', \ldots, \mathbf{Z}_{Pn}')'$ and $\mathbf{Z}_T = \Sigma_{\oplus}\mathbf{Z}_{Ti}$. Hence, in the dispersion matrix $\mathbf{B} = \Sigma_{\oplus}\mathbf{B}_0$, $\mathbf{P} = \Sigma_{\oplus}\mathbf{P}_0$, $\mathbf{G} = \mathbf{A}_n \otimes \mathbf{G}_0$ and $\mathbf{R} = \Sigma_{\oplus}\mathbf{R}_0$, where $\mathbf{A}_n$ is the numerator relationship matrix, $\Sigma_{\oplus}$ denotes direct sum and $\otimes$ represents direct product operation (SEARLE 1982). Since the genetic relationships in our study are limited to half-sib information, it is possible to fit the equivalent half-sib ('sire') model with family rather than tree as the random factor. We keep the tree model notation for the sake of generality.

***Parameterisations of the model***

The expected value of $\mathbf{y}_i$ is $\mathbf{X}_i\,\mathbf{m}$ (Equation 2). This is used to model the average performance of trees as fixed effects, and in this case is the mean at each age. From Equation 3, the dependence of the variance of $\mathbf{y}_i$ on the specification of $\mathbf{B}_0$, $\mathbf{P}_0$, $\mathbf{G}_0$ and $\mathbf{R}_0$ is clear. Since our main interest in the analysis is $\mathbf{G}_0$ thus, following COELLI et al. (1998), we will fit an unstructured error covariance matrix ($\mathbf{R}_0$) while examining various forms for the additive genetic covariance matrix. For the block ($\mathbf{B}_0$) covariance matrix we use an autoregressive correlation structure with separate variances at each time, because it matches our general expectation that the correlation would reduce as the time interval increases. Since there are only 8 blocks, an unstructured form for $\mathbf{B}_0$ would be singular and not estimable. For the plot ($\mathbf{P}_0$) matrix we will primarily use a similar autoregressive correlation structure with heterogeneous variance, but will also report some results from fitting an unstructured form.

We examine the following forms for the additive genetic covariance $\mathbf{G}_0$ (examples of the structures are displayed in Figure 1):

1- Full multivariate model (US). Here $\mathbf{G}_0$ is an unstructured matrix.

2- Banded correlation model (BC) with heterogeneous variances. Here $\mathbf{G}_0 = \mathbf{S}\,\mathbf{C}_{BC}\,\mathbf{S}$ where $\mathbf{S}$ is diagonal matrix of the square roots of the genetic variance components at each age and $\mathbf{C}_{BC}$ is a banded correlation matrix with a specific correlation for each particular age interval.

**Figure 1:** Covariance structures fitted in this study (example using only 4 ages/measures). Different letters (a, b, etc) represent different values of correlation. US: unstructured, BC: banded correlations, AR: autoregressive ($t_j$ is age at measurement j), RR: random regressions expressed as the product $\mathbf{Q}_i\ \Lambda_0\ \mathbf{Q}_i$ ($z_{ij}$ is the $i^{th}$ orthogonal polynomial vector evaluated at age j), and UC: uncorrelated. See text for more detail in the explanation.

$$\begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix}$$

US

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

BC

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

AR

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

RR

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

UC

3- Autoregressive model (AR) with heterogeneous variances. Here $\mathbf{G}_0 = \mathbf{S}\ \mathbf{C}_{AR}\ \mathbf{S}$ where $\mathbf{S}$ is as above and $\mathbf{C}_{AR}$ has an autoregressive correlation structure. We use a power formulation (DIGGLE et al. 1994) which allows for the unequal age intervals, and compare sub-models where the age is expressed on the natural ($\rho^{|k-j|}$, ARnat), square root($\rho^{|\sqrt{k}-\sqrt{j}|}$, ARsqr) and logarithmic($\rho^{|\log(k)-\log(j)|} = \rho^{|\log(k/j)|}$, ARlog) scales, where $\rho$ is a correlation coefficient and j and k are the two ages. Note that when lag (age interval) is expressed in a logarithmic scale the correlation depends upon an age ratio.

4- Random regression model (RR) using orthogonal polynomials. Here $\mathbf{G}_0 = \mathbf{Q}_i\ \Lambda_0\ \mathbf{Q}_i'$, where $\Lambda_0$ is the random regressors covariance matrix and $\mathbf{Q}_i$ has q + 1 columns containing $z_0$, $z_1$, $z_2$, ..., $z_q$ respectively, where q is the order of the polynomial and $z_i$ is the $i^{th}$ orthogonal polynomial vector. Additionally, $\mathbf{a}_i = \mathbf{Q}_i\ \lambda_i$ where $\lambda_i$ are the random regression coefficients.

5- Lastly, an uncorrelated model (UC) (for estimating heritabilities only) is fitted, which is equivalent to a traditional univariate analysis by age. Here all covariances in $\mathbf{G}_0$ (as well as in $\mathbf{R}_0$) are zero.

All models are fitted by restricted maximum likelihood (REML, PATTERSON and THOMPSON 1971) using the average information algorithm (GILMOUR et al. 1995) implemented in ASReml (GILMOUR et al. 1998).

***Model selection***

Adding variance parameters to a model may result in a better fit and hence increase the likelihood value. Several criteria may be used to judge whether additional parameters are making an important contribution to the fit. The Likelihood ratio test formally tests whether the increase is statistically significant. Akaike's Information Criterion and Bayesian Information Criterion (AKAIKE 1974, JONES 1993, CARLIN and LOUIS 1996) penalise the likelihood by the number of independently fitted parameters used in the model. Based on previous experiences in genetic analyses (WADA and KASHIWAGI 1990) we will use Akaike's criterion (AIC), which penalises likelihood values in such a way that any extra parameter must increase the likelihood by at least one unit to be included in the model:

$$AIC = -2\ LogL + 2\ p$$

where –2 LogL is twice the negative log-likelihood value for the model and p is the number of estimated parameters. Smaller values of AIC reflect an overall better fit.

***Genetic parameters***

Estimates of heritability ($h_j^2$) at age j and genetic correlation ($r_{jk}$) between ages j and k are calculated as:

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

$$\hat{r}_{jk} = \frac{\hat{\sigma}_{a_{jk}}}{\sqrt{\hat{\sigma}^2_{a_j} * \hat{\sigma}^2_{a_k}}}$$

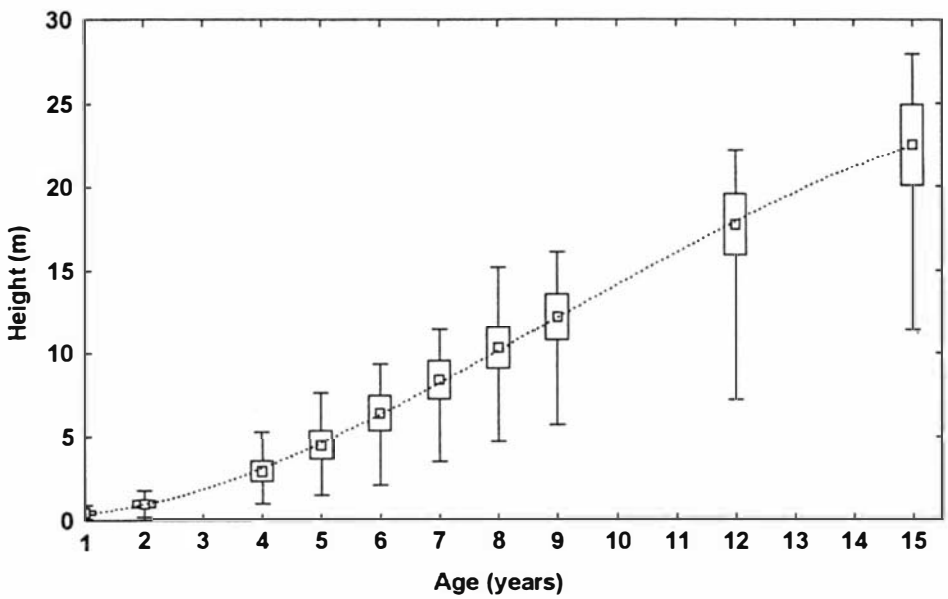
using the appropriate variance and covariance estimates from $\mathbf{G}_0$, $\mathbf{B}_0$, $\mathbf{P}_0$ and $\mathbf{R}_0$. Standard errors of the estimates are calculated by ASReml from the average information matrix, using a standard Taylor series approximation (GILMOUR et al. 1998).

## Results and discussion

### *Exploratory analysis*

Figure 2 shows a box plot of height versus age. The midpoint, box and whiskers represent the mean, mean ± standard deviation and minimum/maximum respectively. This plot depicts an increase of variance with time, which is typical of longitudinal data for growth. It also reveals that the distribution of heights at particular ages is not symmetric after year 5. This asymmetry would be aggravated if a variance stabilising transformation was applied to the data. The figure suggests that a cubic polynomial (broken line) is a reasonable model to form the basis for modelling genetic effects as random regressions.

**Figure 2:** Box plot of height versus age. The midpoint, the box and the whiskers represent the mean, mean ± standard deviation and minimum/maximum respectively. A cubic polynomial (·····) is fitted to the data.




### *Variance models*

The log-likelihood values for a range of models are in Table 2. They range from -3418.2 for the UC model to 7098.9 (7153.2) for the US model with autoregressive (unstructured) plot error structure. Using AIC, the best model (i.e. the one with the lowest value) for both plot error structures is the base model, i.e. the one with no tree effects fitted. Of those with tree effects fitted, the best is the ARlog structure having age

on a natural logarithm basis with AIC = -13961.2. It is followed by the ARsqr, the ARnat, the BC, and the US models (Table 2). The RR model is considered later. The ARnat and ARsqr models have slightly lower heritabilities and slightly higher genetic correlations than the ARlog model, and so their parameters are not included in Figures 3 and 4. In the following discussion we refer to models fitted with an AR plot variance, since not all models would converge to positive definite solution when fitted with a US plot variance model. The genetic parameters with AR plot variance were not much different from those with US plot variance.

**Table 2:** Comparison of multivariate models.

| Model | Number of parameters ($\mathbf{B}_0 + \mathbf{P}_0 + \mathbf{G}_0 + \mathbf{R}_0$) | Log likelihood | AIC[a] |
|---|---|---|---|
| Autoregressive $\mathbf{P}_0$. Base model | 11 + 11 + 0 + 55 = 77 | 7062.51 | -13971.02 |
| Full multivariate (US)[b] | 11 + 11 + 55 + 55 = 132 | 7098.87 | -13933.74 |
| Banded correlations (BC) | 11 + 11 + 21 + 55 = 98 | 7074.94 | -13953.88 |
| Autoregressive (ARnat - age) | 11 + 11 + 11 + 55 = 88 | 7067.04 | -13958.08 |
| Autoregressive (ARsqr - $\sqrt{\text{age}}$) | 11 + 11 + 11 + 55 = 88 | 7067.85 | -13959.70 |
| Autoregressive (ARlog - log(age))[c] | 11 + 11 + 11 + 55 = 88 | 7068.60 | -13961.20 |
| Random regression (RR)[d] | 11 + 11 + 10 + 55 = 87 | 7078.29 | -13982.58 |
| Unstructured $\mathbf{P}_0$. Base model | 11 + 55 + 0 + 55 = 121 | 7124.78 | -14007.56 |
| Full multivariate (US)[e] | 11 + 55 + 55 + 55 = 176 | 7153.17 | -13954.34 |
| Autoregressive (ARlog - log(age))[f] | 11 + 55 + 11 + 55 = 132 | 7127.84 | -13991.68 |
| Uncorrelated (UC) | 10 + 10 + 10 + 10 = 40 | -3418.23 | 6916.46 |

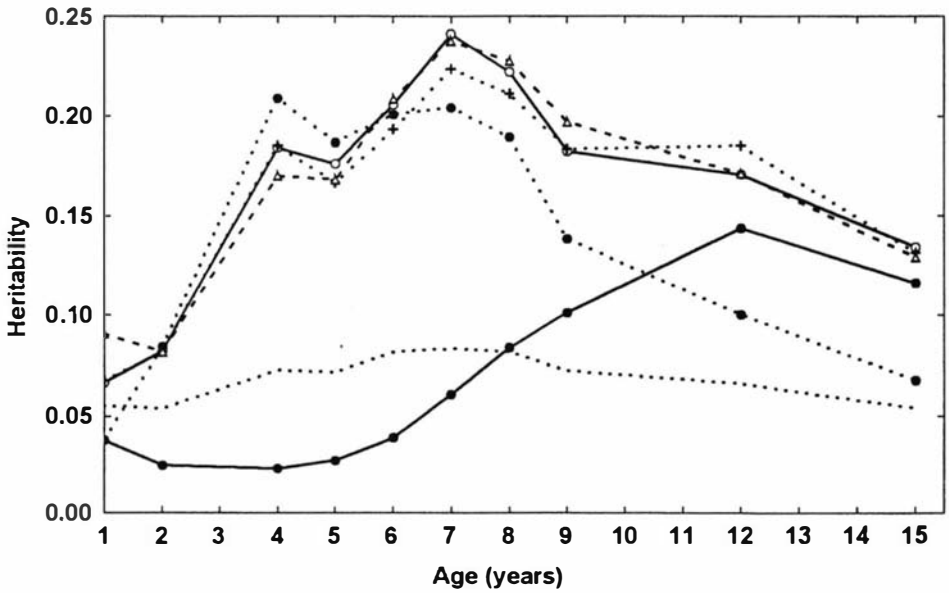
[a] AIC = -2 Log likelihood + 2 number of parameters
[b] $\mathbf{G}_0$ is non positive definite, log-likelihood reduces to 6854.05 after bending.
[c] Best model including $\mathbf{G}_0$ and considering autoregressive $\mathbf{P}_0$.
[d] Reduced rank version: the model converged only by fixing the variance for the quadratic component.
[e] $\mathbf{G}_0$ is non positive definite.
[f] Best model including $\mathbf{G}_0$ and considering unstructured $\mathbf{P}_0$.

**Table 3:** Variance parameters estimated from the autoregressive (ARlog) model with autoregressive plot error: block ($b^2$), plot ($p^2$), additive genetic ($h^2$) and residuals ($e^2$) variances expressed as proportion of phenotypic variance ($\hat{\sigma}_p^2$); correlations among residuals (below diagonal) and genetic correlations (above diagonal).

| Age | $b^2$ | $p^2$ | $h^2$ | $e^2$ | $\hat{\sigma}_p^2$ | Age | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | 12 | 15 |
| 1 | 0.049 | 0.180 | 0.091 | 0.680 | 0.016 | | 0.932 | 0.868 | 0.849 | 0.833 | 0.820 | 0.809 | 0.799 | 0.776 | 0.759 |
| 2 | 0.068 | 0.193 | 0.082 | 0.658 | 0.051 | 0.697 | | 0.932 | 0.911 | 0.894 | 0.880 | 0.868 | 0.858 | 0.833 | 0.814 |
| 4 | 0.057 | 0.178 | 0.170 | 0.596 | 0.380 | 0.607 | 0.758 | | 0.978 | 0.960 | 0.945 | 0.932 | 0.921 | 0.894 | 0.874 |
| 5 | 0.050 | 0.162 | 0.168 | 0.620 | 0.754 | 0.560 | 0.683 | 0.836 | | 0.982 | 0.966 | 0.953 | 0.942 | 0.915 | 0.894 |
| 6 | 0.057 | 0.166 | 0.209 | 0.568 | 1.108 | 0.501 | 0.627 | 0.786 | 0.887 | | 0.984 | 0.971 | 0.960 | 0.932 | 0.911 |
| 7 | 0.060 | 0.141 | 0.238 | 0.561 | 1.309 | 0.469 | 0.581 | 0.731 | 0.831 | 0.916 | | 0.987 | 0.975 | 0.947 | 0.925 |
| 8 | 0.040 | 0.125 | 0.228 | 0.606 | 1.573 | 0.419 | 0.528 | 0.675 | 0.763 | 0.855 | 0.901 | | 0.988 | 0.960 | 0.938 |
| 9 | 0.055 | 0.095 | 0.197 | 0.653 | 2.016 | 0.399 | 0.492 | 0.633 | 0.714 | 0.820 | 0.882 | 0.890 | | 0.971 | 0.949 |
| 12 | 0.026 | 0.067 | 0.171 | 0.737 | 3.399 | 0.361 | 0.412 | 0.533 | 0.620 | 0.719 | 0.779 | 0.812 | 0.832 | | 0.978 |
| 15 | 0.006 | 0.015 | 0.129 | 0.849 | 6.959 | 0.362 | 0.414 | 0.529 | 0.611 | 0.708 | 0.770 | 0.811 | 0.820 | 0.898 | |

The proportions by age for all variance components in the ARlog model are in Table 3. Plot variances are the largest component (18%) in the early years slowly declining in relative magnitude by about 1% per year. The early plot variation might reflect carry-over effects from the nursery. The block variance component is around 5% before declining after age 9. The residual variance is stable at around 60% of phenotypic variance until age 9 increasing to 85% at age 15.

**Figure 3:** Heritability estimates based on fitting the unstructured (US: —o—), banded correlation (BC: ···•···), autoregressive (ARlog: ---Δ---), random regressions (RR: —•—) and uncorrelated (UC: ···+···) models considering autoregressive plot errors, and standard error of the heritability for the autoregressive (·····) model.




Heritability estimates under the ARlog model (Table 3) increase with age from 0.091 at age 1 to 0.238 at age 7 and then decline to 0.129 at age 15. The BC model gave higher heritability before age 8 and lower heritability for later measurements (Figure 3). However, compared with the ARlog model, an increase in the likelihood of 6.3 with 10 extra parameters indicates that the model does not fit the data significantly ($P>0.05$) better. Heritabilities from the US and UC models are very similar to the values from the ARlog model, reaching a maximum of 0.241 and 0.224 respectively (Figure 3). The figure also includes the asymptotic estimate of the standard error of the heritability from

the ARlog model. There is little difference in the standard error of the heritability estimates from the AR (with any time scale), BC, US and UC models.

The heritability differences among these models are quite small until age 6, when the BC model starts underestimating later values. The benefits of a multivariate versus a univariate approach are clearer when covering traits with large differences between genetic and residual correlations, contrasting heritabilities (THOMPSON and MEYER 1986) or there are many missing values for some traits, especially if not missing at random. Here however, the correlations are mostly high, all heritabilities are moderate and only 16% of trees have missing values, particularly at ages 12 and 15.

**Figure 4:** Genetic correlations for different lags between measures for unstructured (US: —o—), banded correlations (BC: ····•····) and autoregressive (ARlog: ---Δ---) models considering autoregressive plot errors.

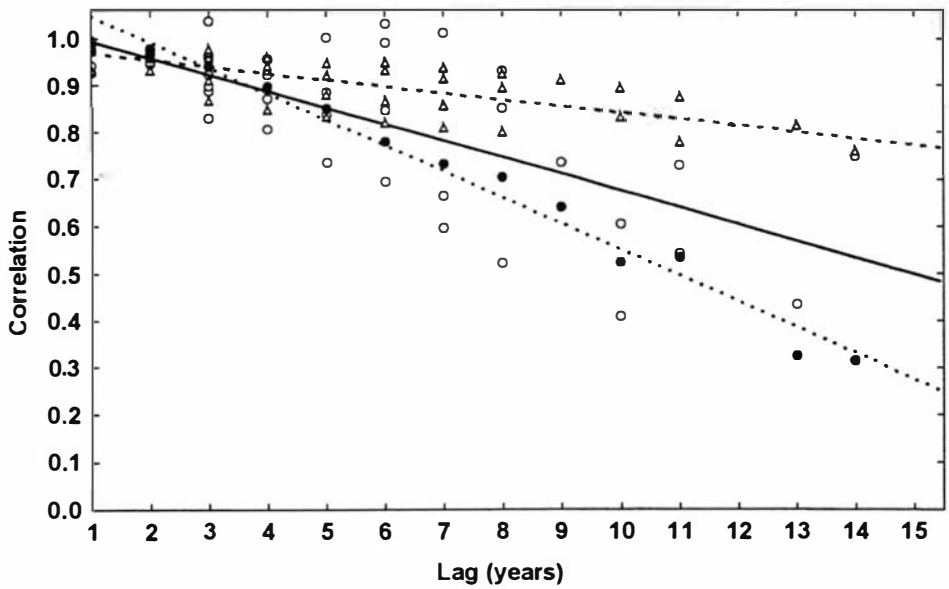



Figure 4 shows the values of genetic correlation plotted against lag for the ARlog, BC and US models. The simple regressions of the correlations on lag shown in Figure 4 had $R^2$ coefficients of 0.58, 0.97, 0.53 for ARlog, BC and US respectively. The correlation between successive ages under the ARlog model increased from 0.932 at age 1 to 0.993 at 14 years of age. The BC model has a single value for each lag whereas the ARlog and

US models have several values for each lag, although the ARnat model has a single value per lag.

The US estimates of $\mathbf{G}_0$ were both non positive definite and with some correlations above 1 (Figure 4). Such a solution is not easily used in practice but commonly occurs due to sampling variation when there are small numbers of families. However, its correlation values help us interpret the other models since they are the least constrained values. The US model gave very high correlations of age 1 with ages 2 to 8 as well as high correlations between successive years; higher than expected under an AR model if the AR parameter was estimated only from more distant ages. Except for the correlations of age 1 noted above, the correlations tended to be lower between younger ages than between higher ages. This is not unexpected when looking at correlations because new growth each year is a decreasing proportion of the height at the beginning of the year.

Genetic correlations for the ARlog model are shown in Table 3. The ARlog model essentially estimates one autocorrelation coefficient and projects the value to higher lags. The projected correlations are higher than the median of the US model values. The high lag correlations from the BC model agree better with the US model values than do the projected ARlog correlations. The US model regression is strongly influenced by the high correlations at all lags associated with age 1. Doing all combinations of bivariate analysis would produce very similar (non positive definite) results. While the US structure puts no constraints on the covariances, the AR and BC structures impose strong constraints, generating more parsimonious models. The variation observed in the US correlations at a given lag represents sampling effects and/or is the result of growth patterns.

The preferred model (ARlog) has an intuitively appealing structure, where the breeding value of tree i observed at 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 ($\alpha_j$), thus $a_{ij} = \rho^{|\log(j)-\log(j-1)|}\, a_{i\,j-1} + \alpha_j$. However, it is necessary to confirm if it is an adequate description of the pattern present in the US correlation structure. The main question posed by the ARlog model is whether the

genetic correlation between ages at higher lag is really so high. As an example, estimates of genetic correlation with the AR model are similar to those obtained for volume by APIOLAZA et al. (1994) until age 9, but overestimate the values for older ages. However, the values are substantially different from the estimates based on clonal material by BURDON et al. (1992). This issue can only be resolved from a much larger experiment. In spite of this, the ARlog model does provide the highest correlations between the ranking of breeding values at different ages and the phenotypic ranking at age 15.

The implications of the high correlation at low lag is that measurement of height in consecutive years contributes little. However, after say five years according to the US and BC models, the correlation has dropped sufficiently to justify remeasurement. The high correlation of age 1 with later ages in the US model (see Figure 4), if real, suggests that early growth is a good predictor of the genetic component of later height. However, the low heritability at age 1 would mitigate against depending on this trait, especially since these could be carry-over nursery effects rather than genetic effects. Further, the high correlations of age 1 with later ages under the US model suggest that whatever causes these early differences, they do persist and are scaled up as the tree grows.

The heritability pattern under the RR model (see Figure 3) did not follow the pattern of the previous models. The correlation pattern is also quite different to the other models with lag 1 correlations ranging from 0.71 at young ages to 0.99 at older ages. One potential explanation was that changes of scale dominated the model, and while the other 4 models had separate variance parameters for each age the RR model had only a function over time. However, a Box-Cox data transformation did not improve much the fitting. The changes in heritability are related to the function used to model $\mathbf{G}_0$. Random regressions model the trajectory of breeding values, which deviate from other fixed and random effects included in the model. Hence a simple polynomial (in this case a third order one) may not be enough to model those deviations. The use of more flexible functions, e.g. higher order polynomials or cubic splines, suggested when there is no previous knowledge about the underlying biological model (e.g. VERBYLA et al. 1997, WHITE et al. 1999), might improve the estimation. Other examples of RR poorly reconstructing the $\mathbf{G}_0$ matrix obtained from a US multivariate analysis are in VAN DER

WERF and SCHAEFFER (1997) and VAN DER WERF et al. (1998). One problem is that the polynomials are by nature highly correlated.

***Number and distribution of measures***

Often the frequency of measurement of a progeny test depends more on budget restrictions than on genetic considerations. Covariance structure modelling may not be possible when only a few measures are available. On the other hand, using 6 measures (1, 4, 6, 9, 12 and 15 years) produced similar results to those obtained using 10 measures (details not shown).

For most longitudinal variance models, equidistant measures are easier to analyse. The presence of unequal intervals involves either the manual specification of the bands (BC) or the use of a distance-based power model (AR). The US model is less likely to converge to a positive definite solution as the number of measures increases or when they are highly correlated. Close measures (as in our data set) result in highly correlated traits increasing the risk that the result might be non positive definite and convergence more difficult (Gilmour 1999).

The process of running the UC analysis (equivalent to univariate analyses by age) was very straightforward from both a modelling and computational perspective. Running the US analysis was less straightforward because of the high correlations between measures (traits). The estimated $\mathbf{G}_0$ was close to singular and the algorithm failed to converge from the naïve starting values initially supplied. Using the results from the BC model as starting values, the algorithm converged to the negative definite solution reported for the US model.

The RR model was quite difficult to fit, also requiring a multistage process to achieve convergence. We first fitted a RR model with intercept and slope, then added the quadratic term and finally the cubic term. The LogL values were 7063.71, 7069.12 and 7078.29 respectively. However, the variance matrix for the final model was negative definite and only converged after fixing the variance for the quadratic component. The correlations among the components were very high despite the definition of $\mathbf{Q}_i$ using orthogonal polynomials. An earlier attempt using starting values derived from the matrix

$\mathbf{Q}_i$' $\mathbf{G}_0$ $\mathbf{Q}_i$ where $\mathbf{G}_0$ is the additive genetic variance matrix from fitting the US model and $\mathbf{Q}_i$ is the 10x4 matrix of orthogonal polynomial coefficients also failed to converge. It was based on the assumption that $\mathbf{G}_0$ is an estimate of the matrix we want to approximate with a structure $\mathbf{Q}_i$ $\Lambda_0$ $\mathbf{Q}_i$' where $\Lambda_0$ is the variance matrix for the random regression coefficients and noting that $\mathbf{Q}_i$' $\mathbf{Q}_i$ is $\mathbf{I}$. That is, if $\mathbf{G}_0 = \mathbf{Q}_i$ $\Lambda_0$ $\mathbf{Q}_i$' then $\mathbf{Q}_i$' $\mathbf{G}_0$ $\mathbf{Q}_i = \mathbf{Q}_i$' $\mathbf{Q}_i$ $\Lambda_0$ $\mathbf{Q}_i$' $\mathbf{Q}_i = \Lambda_0$. We did not try using other polynomials that might change the convergence behaviour.

***Further considerations***

The use of a multivariate approach takes into account non-random missing observations, caused by mortality, thinning or sampling of trees, which can bias parameter estimates (e.g. APIOLAZA et al. 1998). Nonetheless non-positive definite matrices, frequently found in forest genetics literature, may still be an issue. 'Bending' (HAYES and HILL 1981) or other techniques for restricting genetic parameter matrices to the parameter space may still be necessary, especially if the US structure is used and the result is negative definite as here. Where there are large scale effects as in this data set, it may be preferable to bend the correlation matrix rather than the covariance matrix unless the data is transformed to stabilise the variance. The Log likelihood for the bent US model (bent and fixed $\mathbf{G}_0$, other parameters iterated to convergence) reduces to 6854.05, which is lower than the results for the ARlog and BC models, confirming the adequacy of using simplified (co)variance models. The best way to reduce the chance of getting a non-positive definite result is to increase the number of families sampled, reducing the sampling error of the variance components.

An alternative approach to the analysis of highly correlated observations, although beyond the scope of this paper, is the use of canonical transformation. This technique creates independent traits that are analysed separately and the results are transformed back to obtain the full parameter matrix with the original traits (measures) (JENSEN and MAO 1988). It has restrictions on the number of random effects used in the model and the pattern of missing observations (LIN and SMITH 1990, DUCROCQ and BESBES 1993).

## Conclusions

The use of structured covariance matrices for longitudinal data constrains the correlations to a pattern dependent on the form of the model, potentially smoothing the estimates of heritability and genetic correlation. It also facilitates model fitting and convergence of the likelihood-maximisation algorithm. Models that take into account the ordering implicit in successive measures are preferred to the unstructured covariance model when assessing the changes of likelihood relative to the number of parameters. The results presented in this paper suggest that the ARlog model reproduced the results from the US model well enough, while simplifying the analysis; therefore the US covariance structure is probably not the 'best' model for longitudinal data. Equally important is that, assuming AIC is an appropriate model selection criterion, small datasets might not provide enough information to discriminate between some of the models (e.g. RR). On the other hand, if the dataset is appropriate AIC appears to be insensitive to substantial differences of genetic parameters.

The results from this study are from a small sample (45 families with up to 40 trees and 10 longitudinal measures of 1 trait, for a total of 15260 records), but they provide a good starting point for analyses involving larger data sets. Once $\mathbf{G}_0$ and $\mathbf{R}_0$ are estimated, covariance functions can be easily developed (e.g following Kirkpatrick's methodology) allowing for a more detailed study of early selection procedures. Further research might contemplate the use of several measurements to improve early prediction of breeding values (currently under preparation by the authors of this study) and modelling the simultaneous change of other growth traits and wood properties, using multi-trait models (VAN DER WERF et al. 1998).

Although recent literature has emphasised the use of random regression for the analysis of longitudinal data (e.g. JAMROZIK and SCHAEFFER 1997, JAMROZIK et al. 1997, MEYER 1998, VAN DER WERF et al. 1998, WHITE et al. 1999), random regressions are not necessarily suitable for all data sets. Given the potential reduction of number of parameters, other functional relationships (e.g. growth models) that could be used with random regressions within a linear model framework to model $\mathbf{G}_0$ should be studied.

Finally, the 'best' parameterisations can be trait-specific, i.e. different traits may require different structures (e.g. COELLI et al. 1998). Thus it is necessary to identify the most appropriate model for each trait considered in a breeding programme. Nevertheless, simple models like AR seem to be flexible enough to hold in many common tree breeding situations.

## Acknowledgements

This paper arose from discussions with Nuno Borralho and Sue Jarvis during the 6WCGALP, Jan. 1998, Armidale, Australia. Their encouragement is much appreciated. Many thanks to Claudio Balocchi and the breeding team of Bosques Arauco S.A., for measuring the trees over such a long period and kindly providing the data set. Valuable comments by Rowland Burdon, Mark Dieters, Greg Dutkowski, Igor Elsabio, Tore Ericsson and Christian de Veer and by the reviewers Nuno Borralho, Gunnar Jansson, Steen Magnussen, and Per Ståhl greatly improved this paper. Luis Apiolaza was supported with NZODA and NZFRI scholarships during the completion of this study.

## Literature cited

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions of Automatic Control* **19**: 716-723.

APIOLAZA, L.A., WHITE, T.L. and HODGE, G.R. 1994. Análisis genético de *Pinus radiata* en la CMG: estimación de heredabilidad, interacción genotipo-ambiente, correlación edad-edad y correlación entre características en los ensayos de progenie de polinización abierta. School of Forest Resources and Conservation, University of Florida.

APIOLAZA, L.A., BURDON, R.D., and GARRICK, D.J. 1998. Effects of sampling on open-pollinated bivariate progeny tests. P 491-494, Vol. 27 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production, 11-16 January 1998, Armidale, New South Wales, Australia.

BALOCCHI, C.E., BRIDGEWATER, F.E., ZOBEL, B.J., and JAHROMI, S. 1993. Age trends in genetic parameters for tree height in a nonselected population of Loblolly pine. *Forest Science* **39**: 231-251.

BOX, G.E.P. and COX, D.R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society B* **26**: 211-243.

BURDON, R.D. 1977. Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. *Silvae Genetica* **26**: 168-175.

BURDON, R.D., BANNISTER, M.H., and LOW, C.B. 1992. Genetic survey of *Pinus radiata*. 5: between-trait and age-age correlations for growth rate, morphology, and disease resistance. *New Zealand Journal of Forestry Science* **22**: 211-227.

CARLIN, B.P. and LOUIS, T.A. 1996. Bayes and empirical Bayes methods for data analysis. Chapman and Hall, London.

CHI, E.M., and REINSEL, G.C. 1989. Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association* **84**: 452-459.

CNAAN, A., LAIRD, N.M., and SLASOR, P. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* **16**: 2349-2380.

COELLI, K.A., GILMOUR, A.R., and ATKINS, K.D. 1998. Comparison of genetic covariance models for annual measurements of fleece weight and fibre diameter. P 31-34, Vol. 24 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production, 11-16 January 1998, Armidale, New South Wales, Australia.

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

DUCROCQ, V., and BESBES, B. 1993. Solution of multiple trait animal models with missing data on some traits. *Journal of Animal Breeding and Genetics* **110**: 81-92.

FOSTER, G. 1986. Trends in genetic parameters with stand development and their influence on early selection for volume growth in loblolly pine. *Forest Science* **32**: 944-959.

GILMOUR, A.R. 1999. Variance structures available in ASREML. P 416-419, Vol. 13 *in* Proceedings of the Association for the Advancement of Animal Breeding and Genetics, 6-10 April 1999, Dubbo, New South Wales, Australia.

GILMOUR, A.R., THOMPSON, R. and CULLIS, B.R. 1995. Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**: 1440-1450.

GILMOUR, A.R., CULLIS, B.R., WELHAM, S.J., and THOMPSON, R. 1998. ASReml users' manual. New South Wales Agriculture, Orange, Australia.

GREGOIRE, T.G., SCHABENBERGER, O., and BARRET, J.P. 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. *Canadian Journal of Forest Research* **25**: 137-156.

HAND, D., and CROWDER, M. 1996. Practical longitudinal data analysis. Chapman & Hall, London.

HAYES, J.F., and HILL, W.G. 1981. Modification of estimates of parameters in the construction of selection indices ('bending'). *Biometrics* **37**: 483-493.

HODGE, G.R., and WHITE, T.L. 1992. Genetic parameter estimates for growth traits at different ages in slash pine and some implications for breeding. *Silvae Genetica* **41**: 252-262.

JAMROZIK, J., and SCHAEFFER, L.R. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. *Journal of Dairy Science* **80**: 762-770.

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

JENSEN, J., and MAO, I.L. 1988. Transformation algorithms in analysis of single trait and of multitrait models with equal design matrices and one random factor per trait: a review. *Journal of Animal Science* **66**: 2750-2761.

JONES, R.H. 1990. Serial correlation or random subject effects? *Communications of Statistics and Simulation B* **19**:1105-1123.

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

KIRKPATRICK, M., LOFSVOLD, D., and BULMER, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**: 979-993.

KIRKPATRICK, M., HILL, W.G., and THOMPSON, R. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genetical Research* **64**: 57-69.

KREMER, A. 1992. Predictions of age-age correlations of total height based on serial correlations between height increments in maritime pine (*Pinus pinaster* Ait.). *Theoretical and Applied Genetics* **85**: 152-158.

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

LIN, C.Y., and SMITH, S.P. 1990. Transformation of multitrait to unitrait mixed model analysis of data with multiple random effects. *Journal of Dairy Science* **73**: 2494-2502.

LOUIS, T.A. 1988. General methods for analysing repeated measures. *Statistics in Medicine* **7**: 29-45.

MEYER, K. 1998. Modeling 'repeated' records: covariance functions and random regression models to analyse animal breeding data. P 517-520, Vol. 25 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production, 11-16 January 1998, Armidale, New South Wales, Australia.

NEWMAN, D.H., and WILLIAMS, C.G. 1991. The incorporation of risk in optimal selection age determination. *Forest Science* **37**: 1350-1364.

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

QUAAS, R.L., ANDERSON, R.D., and GILMOUR, A.R. 1984. BLUP school handbook. 5-7 February 1984, Animal Genetics and Breeding Unit, University of New England, Australia.

RAO, C.R. 1965. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52**: 447-458.

SCHAEFFER, L.R., and DEKKERS, J.C.M. 1994. Random regressions in animal models for test-day production in dairy cattle. P 443-446, Vol. XVIII *in* Proceedings of the 5th World Congress of Genetics Applied to Livestock Production, August 1994, Guelph, Canada.

SEARLE, S.R. 1982. Matrix algebra useful for statistics. John Wiley and Sons, Inc. New York.

THOMPSON, R., and MEYER, K. 1986. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livestock Production Science* **15**: 299-313.

VAN DER WERF, J.H.J., and SCHAEFFER, L.R. 1997. Random regression in animal breeding. Course notes. 25-28 Jun. 1997, CGIL Guelph, Canada.

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

VERBYLA, A.P., CULLIS, B.R., KENWARD, M.G., and WELHAM, S.J. 1997. The analysis of designed experiments and longitudinal data using smoothing splines. Research Report 97/4, Department of Statistics, The University of Adelaide, Australia.

VONESH, E.F., and CHINCHILLI, V.M. 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker Inc., New York.

WADA, Y., and N. KASHIWAGI. 1990. Selecting statistical models with information statistics. *Journal of Dairy Science* **73**: 3575-3582.

WEI, X. and BORRALHO, N.M.G. 1996. A simple model to describe age trends in heritability in short rotation tree species. P 178-181 *in* DIETERS, M.J., MATHESON, A.C., NIKLES, D.G., HARWOOD, C.E., and WALKER, S.M. (Ed.) Proceedings of Tree Improvement for Sustainable Tropical Forestry, 27 October-1 November 1996, Caloundra, Queensland, Australia.

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

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

# CHAPTER SIX

# OPTIMISING EARLY SELECTION USING LONGITUDINAL DATA

L.A. APIOLAZA[1,2], D.J. GARRICK[1] and R.D. BURDON[2]

[1]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand. [2]New Zealand Forest Research Institute, Private Bag 3020, Rotorua, New Zealand.

SUBMITTED TO SILVAE GENETICA

## Abstract

This study analysed the use of longitudinal data, i.e. repeated assessment of the same individuals at different ages, in the context of early selection. Autoregressive relationships, banded correlations and unstructured ('unsmoothed') matrices were used to model the additive genetic covariance matrix ($\mathbf{G}_0$) for 10 total height measurements of a *Pinus radiata* open-pollinated progeny test. We examined the effects on response to selection of inferred covariance structure, mass versus combined selection, one or multiple assessments, and two breeding-delay intervals. End results are expressed as predicted average gain per year. The patterns of predicted response to selection vary widely between inferred covariance structures. Considering the autoregressive model (based on logarithm of age ratios between assessments) as an example, the effect of combining information from relatives on response to selection is more important (16% to 41% extra gain) than using extra measurements (2% to 25%) when predicting individual breeding values, although the economics of extra gain vs extra assessment costs must be carefully analysed. It is expected that using multiple assessments could be advisable for datasets with lower genetic autocorrelations. An approximate comparison across covariance models showed the autoregressive model to exhibit the best ability to produce 'correct' selections as well as the highest predicted response to selection.

## Introduction

Trees included in breeding programs are often evaluated in progeny trials to predict their genetic value. Results from testing determine the participation of the trees in the breeding population, as well as their use as parents of future plantations (ZOBEL and TALBERT, 1984; WHITE, 1987). The breeding objective includes tree characteristics at harvest age (e.g. volume and wood density); however, progeny tests are assessed at one or more early ages, often less than half the rotation age. The problem of early selection arises with the existence of less than perfect genetic and phenotypic correlations between performance at early assessments and performance at harvest age.

Extending the testing period increases accuracy of selection, i.e. the correlation between predicted and real breeding values, but also increases financial costs and time delays to achieve gain. Optimising response to selection for a given objective involves finding

the appropriate combination of accuracy and evaluation time. Traditionally, this has been achieved by calculating gain for different selection ages using the formula for correlated response to mass or index selection, which includes the heritability at early and mature ages and genetic correlation between the ages (SEARLE, 1965; FALCONER and MACKAY, 1996). The selection age that maximises either a biological criterion (response per year, e.g. LAMBETH, 1980) or an economic criterion (net present value, e.g. NEWMAN and WILLIAMS, 1991) may be chosen.

Genetic and phenotypic covariance structures used for calculating correlated response are often estimated from studies based on a few measurements per tree. Heritabilities are interpolated and extrapolated, perhaps using regression or splines (GWAZE et al., 1997), while genetic correlations are often modelled adapting Lambeth's empirical phenotypic relationship (LAMBETH, 1980), based on either phenotypic or genetic correlations (LAMBETH, 1980; BURDON et al., 1992; GWAZE et al., 1997). WEI and BORRALHO (1996) proposed an alternative model for heritability based on the concept of repeatability. Positive definite additive genetic and phenotypic covariance matrices are not automatically assured by using any of these methods. Concurrently, MAGNUSSEN (1988, 1993) put forward different approaches to early selection, based on the size class distribution of the phenotypes at different ages. However, these latter procedures do not take into account genetic information.

Genetic evaluation tools in forest genetics have undergone a progressive refinement, from evaluation based on family-average (e.g. HATCHER et al. 1981) to the use of Best Linear Prediction (BLP, e.g. WHITE et al., 1987, WHITE and HODGE, 1988) and Best Linear Unbiased Prediction (BLUP, e.g. BORRALHO, 1995, JARVIS et al., 1995). In spite of this, determination of optimum selection time is often based only on response to selection using the most recent measurement of individual performance, even when more assessments were available at the time of analysis. BURDON (1989) suggested the use of longitudinal data, i.e. repeated assessment of the same individuals at different ages, to increase accuracy and therefore response to selection. These assessments can be integrated into a selection index.

This research analyses the implications of using longitudinal data when selecting at an early age. In the course of that we consider the effects of assuming different models for

additive genetic variance, the use of repeated assessments combined in a selection index, and the use of mass and combined selection on the prediction of genetic gain. End results are expressed in terms of average predicted response per year.

## Materials and Methods

### *Dataset*

Genetic and phenotypic covariance matrices for constructing the indexes were estimated from a radiata pine (*Pinus radiata* D. Don) open-pollinated progeny test with 10 assessments at ages 1, 2, 4 to 9, 12, and 15 years from planting. The test included 45 open-pollinated families, planted in 5-tree row-plots within 8 randomised complete blocks. Trees suppressed by early competition were omitted from the analysis, leaving a total of 1526 trees. Observations after mechanical damage to leaders (especially from age 12 onwards) were omitted. Further details are described elsewhere (APIOLAZA et al., 2000).

### *Statistical model*

Considering s assessments on individual i and defining $\mathbf{y}_i = [y_{i1}\ y_{i2} \ldots y_{is}]'$ as the vector of phenotypic observations, the model equation for individual i is:

$$\mathbf{y}_i = \mathbf{X}_i\,\mathbf{m} + \mathbf{Z}_{Bi}\,\mathbf{b} + \mathbf{Z}_{Pi}\,\mathbf{p} + \mathbf{Z}_{Ti}\,\mathbf{a}_i + \mathbf{e}_i$$

where $\mathbf{m} = [m_1\ m_2 \ldots m_s]'$ is the vector of fixed effects (overall mean at each age), $\mathbf{b} = [b_1\ b_2 \ldots b_{bxs}]'$ is the vector of b x s random block effects, $\mathbf{p} = [p_1\ p_2 \ldots p_{pxs}]'$ is the vector of p x s random plot effects, $\mathbf{a}_i = [a_{i1}\ a_{i2} \ldots a_{is}]'$ is the vector of additive genetic values, $\mathbf{e}_i = [e_{i1}\ e_{i2} \ldots e_{is}]'$ is the vector of random residuals, and $\mathbf{X}_i$, $\mathbf{Z}_{Bi}$, $\mathbf{Z}_{Pi}$ and $\mathbf{Z}_{Ti}$ are the incidence matrices for fixed, block, plot and additive genetic effects respectively. If we think of a non-inbred individual with all measurements (i.e. no missing observations) the dispersion matrices are:

$$\text{Var}[\mathbf{b}] = \mathbf{B}_0,\ \text{Var}[\mathbf{p}] = \mathbf{P}_0,\ \text{Var}[\mathbf{a}_i] = \mathbf{G}_0 \text{ and } \text{Var}[\mathbf{e}_i] = \mathbf{R}_0$$

with typical elements $\sigma_{b_{jk}}$, $\sigma_{p_{jk}}$, $\sigma_{a_{jk}}$, $\sigma_{e_{jk}}$ respectively. For individuals with missing observations (and the vector $\mathbf{y}_i$ reduced accordingly) the corresponding rows and columns from $\mathbf{R}_0$ are omited.

The multivariate model equation for the N individuals in the progeny test is:

$$\mathbf{y} = \mathbf{X}\,\mathbf{m} + \mathbf{Z}_B\,\mathbf{b} + \mathbf{Z}_P\,\mathbf{p} + \mathbf{Z}_T\,\mathbf{a} + \mathbf{e}$$

where $\mathbf{y} = [\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N']'$, $\mathbf{a} = [\mathbf{a}_1', \mathbf{a}_2', \ldots, \mathbf{a}_N']'$ and $\mathbf{e} = [\mathbf{e}_1', \mathbf{e}_2', \ldots, \mathbf{e}_N']'$. In addition $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_N']'$, $\mathbf{Z}_B = [\mathbf{Z}_{B1}', \mathbf{Z}_{B2}', \ldots, \mathbf{Z}_{BN}']'$, $\mathbf{Z}_P = [\mathbf{Z}_{P1}', \mathbf{Z}_{P2}', \ldots, \mathbf{Z}_{PN}']'$ and $\mathbf{Z}_T = \Sigma_{\oplus}\mathbf{Z}_{Ti}$, where $\Sigma_{\oplus}$ denotes direct sum (SEARLE, 1982).

The expected value and dispersion matrices considering a multivariate normal distribution and zero covariance between random factors (blocks, plots and trees) are:

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

$$Var[\mathbf{b}] = \mathbf{B} = \Sigma_{\oplus}\mathbf{B}_0,\ Var[\mathbf{p}] = \mathbf{P} = \Sigma_{\oplus}\mathbf{P}_0,\ Var[\mathbf{a}] = \mathbf{G} = \mathbf{A}_N \otimes \mathbf{G}_0 \text{ and } Var[\mathbf{e}] = \mathbf{R} = \Sigma_{\oplus}\mathbf{R}_0$$

$$\text{thus } Var[\mathbf{y}] = \mathbf{Z}_B\,\mathbf{B}\,\mathbf{Z}_B' + \mathbf{Z}_P\,\mathbf{P}\,\mathbf{Z}_P' + \mathbf{Z}_T\,\mathbf{G}\,\mathbf{Z}_T' + \mathbf{R}$$

where $\mathbf{A}_N$ is the numerator relationship matrix (HENDERSON, 1984) and $\otimes$ denotes direct product (SEARLE, 1982).

Best Linear Unbiased Prediction of the breeding values ($\mathbf{a}$) were calculated using Henderson's mixed model equations (HENDERSON, 1984) and assumed values of covariance components. Values for all covariance components were obtained by Restricted Maximum Likelihood (REML) using ASReml (GILMOUR et al. 1998).

**Figure 1:** Example of the unstructured (US), banded correlations (BC) and autoregressive with time on a natural logarithm scale (ARlog) models. Correlations with the same letter represent the same value. The BC model assumes similar correlations for measurements with equal time between expressions.

$$\mathbf{C}_{US} = \begin{bmatrix} 1 & a & b & c \\ a & 1 & d & e \\ b & d & 1 & f \\ c & e & f & 1 \end{bmatrix} \quad \mathbf{C}_{BC} = \begin{bmatrix} 1 & g & h & i \\ g & 1 & g & h \\ h & g & 1 & g \\ i & h & g & 1 \end{bmatrix} \quad \mathbf{C}_{AR\log} = \begin{bmatrix} 1 & r^{|\log(2/1)|} & r^{|\log(3/1)|} & r^{|\log(4/1)|} \\ r^{|\log(2/1)|} & 1 & r^{|\log(3/2)|} & r^{|\log(4/2)|} \\ r^{|\log(3/1)|} & r^{|\log(3/2)|} & 1 & r^{|\log(4/3)|} \\ r^{|\log(4/1)|} & r^{|\log(4/2)|} & r^{|\log(4/3)|} & 1 \end{bmatrix}$$

APIOLAZA et al. (2000) analysed the dataset comparing five models for describing the additive genetic covariance matrix $\mathbf{G}_0$: unstructured (US), autoregressive with time in a natural logarithm time scale (ARlog), banded correlations (BC), random regressions (RR), and uncorrelated (UC). The UC model is equivalent to independent univariate analyses and does not provide direct estimates for genetic correlations between ages (i.e. assuming zero between-trait genetic correlations). The RR model did not converge;

thus its estimates of genetic parameters are not reliable. Therefore we calculated response to selection from only three models: US, BC and ARlog.

**Table 1:** Genetic parameters for unstructured (US), banded correlations (BC) and autoregressive with time on a natural logarithm scale (ARlog) models. Additive genetic ($\sigma_a^2$) and phenotypic variances ($\sigma_y^2$), additive genetic covariances ($\sigma_{a_{jk}}$, above diagonal) and phenotypic covariances ($\sigma_{y_{jk}}$, below diagonal).

| | | | Age (years) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | $\sigma_a^2$ | $\sigma_p^2$ | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | 12 | 15 |
| US | | | | | | | | | | | | |
| 1 | 0.001 | 0.016 | | 0.002 | 0.007 | 0.011 | 0.016 | 0.019 | 0.019 | 0.018 | 0.018 | 0.024 |
| 2 | 0.004 | 0.051 | 0.022 | | 0.016 | 0.023 | 0.029 | 0.032 | 0.033 | 0.026 | 0.020 | 0.027 |
| 4 | 0.070 | 0.379 | 0.054 | 0.113 | | 0.093 | 0.120 | 0.132 | 0.136 | 0.118 | 0.105 | 0.139 |
| 5 | 0.132 | 0.752 | 0.071 | 0.147 | 0.473 | | 0.169 | 0.194 | 0.193 | 0.178 | 0.166 | 0.213 |
| 6 | 0.227 | 1.105 | 0.080 | 0.167 | 0.548 | 0.837 | | 0.264 | 0.273 | 0.266 | 0.253 | 0.339 |
| 7 | 0.315 | 1.306 | 0.083 | 0.170 | 0.563 | 0.863 | 1.127 | | 0.330 | 0.333 | 0.358 | 0.463 |
| 8 | 0.349 | 1.570 | 0.083 | 0.173 | 0.579 | 0.887 | 1.170 | 1.331 | | 0.349 | 0.394 | 0.511 |
| 9 | 0.367 | 2.014 | 0.089 | 0.183 | 0.616 | 0.946 | 1.270 | 1.464 | 1.633 | | 0.444 | 0.581 |
| 12 | 0.580 | 3.402 | 0.102 | 0.199 | 0.678 | 1.065 | 1.439 | 1.678 | 1.924 | 2.241 | | 0.764 |
| 15 | 0.936 | 6.968 | 0.130 | 0.256 | 0.878 | 1.399 | 1.893 | 2.224 | 2.597 | 3.014 | 4.354 | |
| BC | | | | | | | | | | | | |
| 1 | 0.001 | 0.016 | | 0.002 | 0.007 | 0.008 | 0.010 | 0.010 | 0.010 | 0.009 | 0.008 | 0.005 |
| 2 | 0.004 | 0.052 | 0.022 | | 0.018 | 0.023 | 0.028 | 0.029 | 0.028 | 0.025 | 0.020 | 0.015 |
| 4 | 0.080 | 0.382 | 0.054 | 0.114 | | 0.105 | 0.130 | 0.138 | 0.138 | 0.127 | 0.115 | 0.103 |
| 5 | 0.142 | 0.759 | 0.071 | 0.148 | 0.478 | | 0.175 | 0.190 | 0.193 | 0.178 | 0.160 | 0.135 |
| 6 | 0.223 | 1.112 | 0.080 | 0.168 | 0.553 | 0.844 | | 0.241 | 0.251 | 0.234 | 0.214 | 0.207 |
| 7 | 0.268 | 1.311 | 0.082 | 0.171 | 0.569 | 0.871 | 1.134 | | 0.279 | 0.266 | 0.256 | 0.249 |
| 8 | 0.298 | 1.572 | 0.083 | 0.174 | 0.585 | 0.894 | 1.176 | 1.335 | | 0.284 | 0.284 | 0.273 |
| 9 | 0.278 | 2.008 | 0.089 | 0.185 | 0.623 | 0.954 | 1.274 | 1.464 | 1.632 | | 0.288 | 0.281 |
| 12 | 0.337 | 3.368 | 0.101 | 0.201 | 0.685 | 1.073 | 1.443 | 1.673 | 1.916 | 2.223 | | 0.373 |
| 15 | 0.468 | 6.900 | 0.129 | 0.256 | 0.877 | 1.393 | 1.880 | 2.197 | 2.564 | 2.970 | 4.296 | |
| Arlog | | | | | | | | | | | | |
| 1 | 0.002 | 0.016 | | 0.002 | 0.008 | 0.012 | 0.015 | 0.018 | 0.019 | 0.019 | 0.023 | 0.028 |
| 2 | 0.004 | 0.051 | 0.022 | | 0.015 | 0.021 | 0.028 | 0.032 | 0.034 | 0.035 | 0.041 | 0.050 |
| 4 | 0.064 | 0.380 | 0.055 | 0.115 | | 0.088 | 0.117 | 0.133 | 0.141 | 0.148 | 0.174 | 0.211 |
| 5 | 0.126 | 0.754 | 0.072 | 0.149 | 0.474 | | 0.168 | 0.192 | 0.203 | 0.212 | 0.250 | 0.304 |
| 6 | 0.231 | 1.108 | 0.082 | 0.170 | 0.549 | 0.839 | | 0.264 | 0.280 | 0.293 | 0.344 | 0.419 |
| 7 | 0.312 | 1.309 | 0.084 | 0.173 | 0.565 | 0.865 | 1.129 | | 0.331 | 0.345 | 0.406 | 0.494 |
| 8 | 0.360 | 1.573 | 0.084 | 0.176 | 0.583 | 0.891 | 1.173 | 1.333 | | 0.376 | 0.443 | 0.538 |
| 9 | 0.402 | 2.016 | 0.090 | 0.187 | 0.622 | 0.953 | 1.274 | 1.465 | 1.636 | | 0.473 | 0.575 |
| 12 | 0.590 | 3.399 | 0.105 | 0.208 | 0.702 | 1.097 | 1.476 | 1.709 | 1.954 | 2.269 | | 0.718 |
| 15 | 0.913 | 6.959 | 0.134 | 0.266 | 0.906 | 1.435 | 1.933 | 2.258 | 2.630 | 3.046 | 4.366 | |

Expressing $\mathbf{G}_0 = \mathbf{S}\ \mathbf{C}\ \mathbf{S}$, where **S** is a diagonal matrix with elements equal to the standard deviations for each assessment and **C** the correlation matrix between assessments, these covariance models can be seen as using **C** with different sets of restrictions (see Figure 1). In the US model $\mathbf{C}_{US}$ has no restrictions except for being positive definite. The BC model assumes similar correlations for measurements with

equal time between expressions, creating bands of identical correlations in $\mathbf{C}_{BC}$. The autoregressive model assumes that in $\mathbf{C}_{ARlog}$ the correlation between two assessments at ages j and k has form $r^{|\log(k)-\log(j)|} = r^{|\log(k/j)|}$. Additive genetic and phenotypic covariance parameters for these three models are presented in Table 1 (adapted from APIOLAZA et al. 2000).

ARlog was selected as the best model from a penalised-likelihood standpoint (i.e. the log-likelihood value LogL penalised, in this case, according to the number p of independent parameters of the model), using Akaike's Information Criterion (AIC, WADA and KASHIWAGI 1990):

$$AIC = -2\ LogL + 2\ p$$

The best model using this criterion has the lowest value of AIC. A detailed description of the models and model selection criteria can be found in APIOLAZA and GARRICK (2000).

***Predicted response to selection***

Consider the breeding objective (H) or aggregate genotype as a linear combination of n additive genetic values $\mathbf{a} = [a_1\ a_2\ \ldots\ a_n]'$ weighted by their relative economic importance $\mathbf{v}' = [v_1\ v_2\ \ldots\ v_n]$ where all elements of $\mathbf{v}$ are different from zero:

$$H = \mathbf{v}'\ \mathbf{a}$$

The selection index used to predict the aggregate genotype is:

$$I = \mathbf{c}'\ \mathbf{y}^*$$

where $\mathbf{c}' = [c_1\ c_2\ \ldots\ c_m]$ is the vector of index weights, calculated using economic and genetic information to maximise the correlation between H and I, and $\mathbf{y}^* = [y_1^*\ y_2^*\ \ldots\ y_m^*]'$ is the vector of phenotypic assessments on the trees adjusted for fixed effects (i.e. $\mathbf{y} - \mathbf{X}\ \mathbf{m}$ for BLP or, in our case, $\mathbf{y} - \mathbf{X}\ \hat{\mathbf{m}}$ for BLUP). Index weights are calculated using (HAZEL, 1943):

$$\mathbf{c} = \mathbf{P}^{-1}\ \mathbf{G}\ \mathbf{v}$$

where $\mathbf{P}$ and $\mathbf{G}$ represent the phenotypic and additive genetic covariance matrices for the traits. These weights maximise the accuracy of selection, i.e. the correlation between I and H ($r_{IH}$) when the fixed effects, $\mathbf{P}$ and $\mathbf{G}$ are known.

Predicted response to selection ($\Delta G$), considering one generation, is a function of accuracy of selection ($r_{IH}$), the variance of the selection target ($\sigma_H^2$) and the intensity of selection (i, related to the proportion of trees selected):

$$\Delta G = i\, r_{IH}\, \sigma_H$$

H can be partitioned in such a way that $H = a_1 I_1 + a_2 I_2 + \ldots + a_n I_n$, where each index estimates the breeding value for a different trait. VILLANUEVA et al. (1993) extended this concept partitioning matrices **P** and **G** to facilitate generalisation of the estimation process, where each submatrix corresponds to a trait. The more general case, for m selection criteria (characters and/or measurements with typical elements j and k) and n traits in the aggregate genotype (with typical element q) is:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1m} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{m1} & \mathbf{P}_{m2} & \cdots & \mathbf{P}_{mm} \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_{11} & \mathbf{g}_{12} & \cdots & \mathbf{g}_{1n} \\ \mathbf{g}_{21} & \mathbf{g}_{22} & \cdots & \mathbf{g}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{m1} & \mathbf{g}_{m2} & \cdots & \mathbf{g}_{mn} \end{bmatrix}$$

$\mathbf{P}_{jk}$ refers to the phenotypic covariance matrix between selection criteria j and k, and $\mathbf{g}_{jq}$ represents the vector of covariances between the selection criteria j and breeding value for trait q. Breeding values on H can be estimated with different selection indices (selection criteria) representing different selection schemes:

- Mass selection, considering only the own record of individual i for trait j ($y_{ij}^*$), representing the simplest index:

$$I_{mass_i} = c_{1j}\, y_{ij}^*$$

- Combined selection includes individual and family-average information for trait j, whereby:

$$I_{comb_i} = c_{1j}\, y_{ij}^* + c_{2j}\, \bar{y}_{.j}^*$$

where $\bar{y}^*_{.j}$ is the phenotypic average (adjusted for fixed effects) for the half-sib family, including the assessment on $y_{ij}$*, on trait/measurement j, and $c_{1j}$ and $c_{2j}$ are index coefficients for individual and family information respectively.

This notation is fairly general, allowing for any number of selection criteria and traits in the breeding objective, and can be readily extended to other types of relatives (see KERR 1998 for examples based on full-sib mating over a number of generations). Consider now an objective including a single breeding value of the individual (e.g. height at 15 years) and m phenotypic assessments at earlier ages (e.g. height at age ≤ 15 years). **P** and **G** for a given selection method contain m x m submatrices and m subvectors respectively. The submatrices and subvectors are:

- Mass selection (in this case $\mathbf{P}_{jk}$ and $\mathbf{g}_{jq}$ have dimension 1x1, i.e. they are scalars):

$$\mathbf{P}_{jk} = \mathrm{Cov}(y_{ij}^*, y_{ik}^*) = \sigma_{y_{jk}}$$

$$\mathbf{g}_{jq} = \mathrm{Cov}(y_{ij}^*, a_{iq}) = \mathrm{Cov}(a_{ij}, a_{iq}) = \sigma_{a_{jq}}$$

- Combined selection (individual - $y_{ij}^*$ - and average of half-sib family - $\bar{y}^*_{.j}$ - with t individuals):

$$\mathbf{P}_{jk} = \begin{bmatrix} \mathrm{Cov}(y^*_{ij}, y^*_{ik}) & \mathrm{Cov}(y^*_{ij}, \bar{y}^*_{.k}) \\ \mathrm{Cov}(\bar{y}^*_{.j}, y^*_{ik}) & \mathrm{Cov}(\bar{y}^*_{.j}, \bar{y}^*_{.k}) \end{bmatrix}$$

$$\mathbf{g}_{jq} = \begin{bmatrix} \mathrm{Cov}(y^*_{ij}, a_{iq}) \\ \mathrm{Cov}(\bar{y}^*_{.j}, a_{iq}) \end{bmatrix}$$

where:

$$\mathrm{Cov}(\bar{y}^*_{.j}, y^*_{ik}) = \mathrm{Cov}(y^*_{ij}, \bar{y}^*_{.k}) = \mathrm{Cov}(\bar{y}^*_{.j}, \bar{y}^*_{.k}) = \left(\sigma_{p_{jk}} + 0.25\,(t-1)\sigma_{a_{jk}}\right)/t$$

$$\mathrm{Cov}(y^*_{ij}, a_{iq}) = \sigma_{a_{jq}}$$

$$\mathrm{Cov}(\bar{y}^*_{.j}, a_{il}) = \frac{\sigma_{a_{jq}}}{2}$$

and the phenotypic covariance is:

$$\sigma_{y_{jk}} = \sigma_{b_{jk}} + \sigma_{p_{jk}} + \sigma_{a_{jk}} + \sigma_{e_{jk}}$$

Further generalisation to an objective with n traits would involve m x n subvectors to define **G**.

***Number and timing of measurements***

We consider from one to three assessments only for progeny tests, owing to economic and practical considerations, to predict height performance at 15 years. Thus **P** contains between 1 x 1 and 3 x 3 submatrices, **G** contains between 1 and 3 subvectors, and the breeding objective considers only 1 trait. We generate all combinations for 1, 2 and 3 measurements out of the 10 assessments used by APIOLAZA et al. (1999) and calculate predicted response to selection for each combination. We do not interpolate intermediate ages of assessments, but use only years actually assessed. For all calculations we assume 200 families and 40 trees per family using forwards selection, i.e. selection of the progeny rather than of the parents. The selection is 200 out of 8000, for an intensity i of 2.338 (FALCONER and MACKAY 1996).

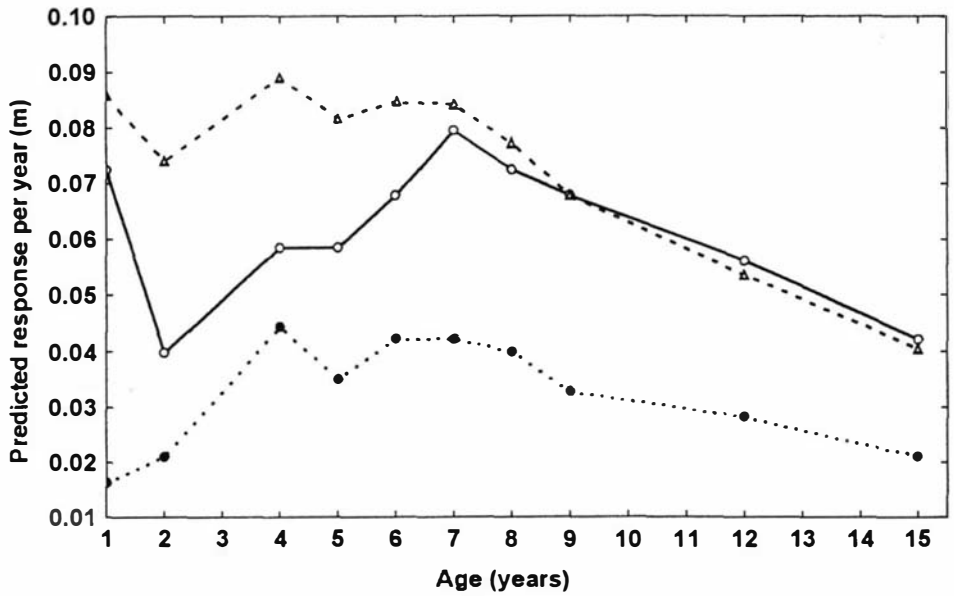
The best option is chosen based on response per unit of time, i.e. response divided by generation interval, ignoring measurement costs. Generation interval is calculated as selection delay (age of the latest measurement) plus breeding delay (time between selection and propagation of sufficient offspring for planting). Two levels for breeding delay are considered: 5 and 8 years.

## Results and discussion

APIOLAZA et al. (2000) determined that, based upon AIC value, the ARlog model was the best for the data set analysed. Nevertheless, they also pointed out that small differences in statistical model selection criterion could conceal large differences in genetic parameters. Given that the dataset available comprises only one generation, it is not possible to compare the covariance structures in terms of empirical gain, but only in terms of predicted response to selection based upon estimates of genetic parameters. Consequently, at each time one of the models was assumed as the 'true' one and response to selection calculated accordingly. Because of this, results are not directly comparable across covariance models.

The pattern of predicted response to selection, considering any number of measurements, vary widely among covariance structures. Both US and ARlog models achieve similar maximum gain per year, but with three years of difference in timing (4 v/s 7 years, Figure 2). The ARlog model consistently achieves higher gains when selecting under age 9 years. Predicted response for the ARlog model tends to be dominated by the high level of autocorrelation, while in the US model seems to follow the trend for heritability of height. The predicted response to selection of the US model based on early measurements seems to fluctuate more erratically. APIOLAZA et al. (2000), suggest the use of a much larger experiment to obtain more reliable estimates of the genetic correlations. Predicted response from the BC model is far lower than with the other two models but follows a trend similar to the ARlog model. In the former model heritability and genetic correlation estimates are most of the time smaller than in the ARlog and US models.

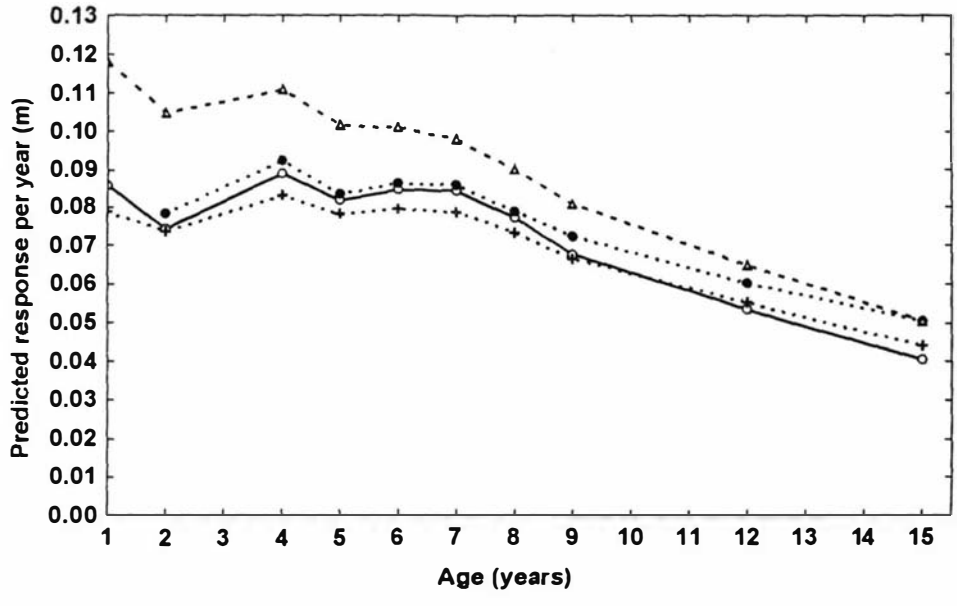
**Figure 2:** Predicted response per year to forwards mass selection for different covariance structures: unstructured (US, —o—), banded correlations (BC, ···•···) and autoregressive with with time on a natural logarithm scale (ARlog, ---Δ---) considering 5 years breeding delay.




The trends of the effects of additional measurements and extra information from relatives were similar for all covariance models. We will use the ARlog model for illustration purposes. The effect of integrating information from relatives in the genetic

evaluation is far greater than that of including extra measurements when predicting the breeding values (see Figure 3), with little or no extra cost (except those related to the use of more sophisticated analyses). Results for multiple assessments presented in Figure 3 correspond to the best combination of two ages lower or equal to the age reported (Table 2). From age 1 to 8 years using an extra measurement increases predicted response between 2 and 5%, and only from age 9 onwards predicted response increases from 7% at 9 years to 25% at 15 years (Figure 3). A similar trend is observed for combined selection with 2 measurements (results not presented). In spite of this, the optimum selection time does not change from age 4 years, when the additional predicted response is 4%. The results from including a third measurement are not presented, because the gain is marginal.

**Figure 3:** Predicted response per year to forward selection for the autoregressive model with time on a natural logarithm scale (ARlog) considering: single assessment-mass selection (—o—), two assessments-mass selection (···•···) and single assessment-combined selection (---Δ---) for 5 years of breeding delay. Response considering single assessment-combined selection for 8 years of breeding delay (···+···).




The use of family information increased predicted response to forwards selection by 16% to 41%, especially at early ages. Additionally, the optimum selection times reduces from 4 years to 1 year. The reduction of selection age also applies to the US

model, where optimum selection is at one year. Selection time is not affected in the BC model.

The only case when combined selection is inferior to mass selection using two assessments (at ages 7 and 15) is for predicted response selecting at age 15. This is caused by the low accuracy at age 15 (low heritability) compared to an index that integrates information from that year with information from age 7 (the age of the highest heritability).

**Table 2:** Predicted response to selection and best combinations of assessments for the autoregressive model with time on a natural logarithm scale (ARlog), for mass selection with 1 and 2 measurements.

| Age of selection (years) | Mass selection, 1 measurement | Mass selection, 2 measurements | |
|---|---|---|---|
| | Response per year (m) | Response per year (m) | Combination of measurements (ages) |
| 1 | 0.086 | | |
| 2 | 0.074 | 0.078 | [1 2] |
| 4 | 0.089 | 0.092 | [2 4] |
| 5 | 0.082 | 0.083 | [4 5] |
| 6 | 0.085 | 0.086 | [2 6] |
| 7 | 0.084 | 0.086 | [2 7] |
| 8 | 0.077 | 0.079 | [7 8] |
| 9 | 0.068 | 0.072 | [7 9] |
| 12 | 0.054 | 0.060 | [7 12] |
| 15 | 0.040 | 0.051 | [7 15] |

When considering selection at very early ages (e.g. 1 year) a breeding delay of 5 years could be far too optimistic, given current biological constraints. Breeding delay includes the time needed for flowering, the delay between flowering and seed production, and time for multiplication. While the last two are independent from selection age, the first one is probably interdependent with age during the first 5 years. Therefore, the use of a

uniform breeding delay for all selection ages should be considered as a simplifying assumption. If the first component of breeding delay is addressed with more detail selection ages would tend to be pushed forward. As expected, the effect of increasing breeding delay (from 5 to 8) is larger for early selection ages than for late selections (Figure 3). The reduction of predicted per-year response resulting from the increased breeding delay, using single-assessment combined selection, ranges from 33% at age 1 to 13% at age 15 (Figure 3). This trend of reductions is close to linear ($R^2$=0.95) and very similar for all covariance structures.

## Final considerations

Although including more assessments increases predicted response to selection, especially after age 8, the extra response does not match the gain attained using a single assessment with combined selection. Nevertheless, it is expected that using multiple assessments could be advisable for datasets with lower genetic autocorrelations or strong age-age environmental correlations. A decision on single versus multiple measurements for selection should take into account the gain in response (weighted by the number of hectares deployed with material from the breeding program) versus the costs of extra measurements. Three further considerations are: that several measurements might reduce optimal generation interval (increasing accuracy at a given age and ascertaining the stability of rankings, making earlier selection more appealing), that selection for seed orchards can be continually updated and could make use of additional measurements (e.g. last three combinations of Table 2), and that measuring costs often increase with age (e.g. NEWMAN and WILLIAMS 1991).

We anticipate that reducing breeding delay (through overcoming biological constraints upon age of flowering) would drive optimum selection to earlier ages, because the denominator of response per unit of time would be dominated by selection age. Reducing the generation interval from 9 years (i.e. selection at 4 years) to 6 years (i.e. selecting at 1 year) most probably will affect profit when considering net discounted value.

It seems that it is still possible in tree breeding to obtain additional gains through the use of more sophisticated methods of genetic evaluation, without resorting to extra

assessments. Dealing with more than one generation and/or more complex crossing designs (e.g. controlled pollination) will imply using selection criteria including the estimated breeding values of the parents. However if selection operates through several generations other factors must be considered: gametic linkage phase disequilibrium (Bulmer effect), genetic drift, mutational variance and effective population size (WEI et al. 1996). An extra advantage of considering data over several generations is the opportunity to determine the best covariance model on terms of realised response to selection, rather than on predicted values.

An important aspect, but beyond the scope of this paper, is the consideration of the risk involved in early selection. Deviations from predicted gain (either overestimation or underestimation) can potentially alter both the selection age(s) and the economic results of a breeding program; and the variance of the response should be taken into account. Risk may arise, among other reasons, because of low accuracy of prediction, traits not being expressed at selection age (e.g. effect of *Cyclaneusma* needle cast, cf. BURDON 1989), differences between performance at final assessment (15 years) and at rotation age (20 years), and the effect of faster reduction of effective population size due to more frequent generation turnover. Several approaches to deal with risk have been proposed in tree and animal breeding literature including simulation of predicted gain using stochastic sampling of genetic correlations (e.g. NEWMAN and WILLIAMS 1991, MAGNUSSEN and YANCHUK 1993), quadratic programming (SCHNEEBERGER et al. 1982) and Bayesian decision theory (e.g. WOOLLIAMS and MEUWISSEN 1993). A comprehensive risk analysis will probably need to consider the effects of early selection on variance of predicted response across several generations.

It is appropriate to emphasise that the results presented in this research relate to a small number of families growing in one site. Therefore they should not be considered as the 'standard' results for radiata pine. It will be necessary to extend the analyses of longitudinal data to datasets including multiple sites and more families, to be sure of the reliability of estimates of genetic parameters. Accordingly, this study provides an illustration of methodology rather than definitive guidelines for early selection.

## Acknowledgements

Luis Apiolaza was funded with New Zealand Overseas Development Assistance and New Zealand Forest Research Institute scholarships during the development of this study. Thanks are due to Claudio Balocchi and the breeding team of Bosques Arauco S.A., who kindly provided the dataset used in this study.

## Literature cited

APIOLAZA, L.A. and GARRICK, D.J. 2000. Analysis of longitudinal data from progeny tests: some multivariate approaches. *Forest Science* (in press).

APIOLAZA, L.A., GILMOUR, A.R. and GARRICK, D.J. 2000. Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. *Canadian Journal of Forest Research* **30**: 645-654.

BORRALHO, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. P. 141-145 *in* POTTS, B.M., BORRALHO, N.M.G., REID, J.B., CROMER, R.N., TIBBITS, W.N., and RAYMOND, C.A. "Eucalypts plantations: improving fibre yield and quality". Proceedings of CRC-IUFRO Conference, 19-24 February, Hobart, Tasmania, Australia.

BURDON, R.D. 1989. Early selection in tree breeding: principles for applying index selection and inferring input parameters. *Canadian Journal of Forest Research* **19**: 499-504.

BURDON, R.D., BANNISTER, M.H. and LOW, C.B. 1992. Genetic survey of *Pinus radiata*. 5: between-trait and age-age correlations for growth rate, morphology, and disease resistance. *New Zealand Journal of Forest Science* **22**: 211-227.

FALCONER, D.S. and MACKAY, T.F.C. 1996. Introduction to quantitative genetics. Longman Group Ltd., UK.

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

GWAZE, D.P., WOOLLIAMS, J.A. and KANOWSKI, P.J. 1997. Optimum selection age for height in *Pinus taeda* L. in Zimbabwe. *Silvae Genetica* **46**: 358-365.

HATCHER, A.V., BRIDGWATER, F.E. and WEIR, R.J. 1981. Performance level – standardized score for progeny test performance. *Silvae Genetica* **30**: 184-187.

HAZEL, L.N. 1943. The genetic basis for constructing selection indexes. *Genetics* **28**: 476-490.

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

JARVIS, S.F., BORRALHO, N.M.G. and POTTS, B.M. 1995. Implementation of a multivariate BLUP model for genetic evaluation. P 212-216 *in* POTTS, B.M., BORRALHO, N.M.G., REID, J.B., CROMER, R.N., TIBBITS, W.N., and RAYMOND, C.A. "Eucalypts plantations: improving fibre yield and quality". Proceedings of CRC-IUFRO Conference, 19-24 February, Hobart, Tasmania, Australia.

KERR, R.J. 1998. Asymptotic rates of response from forest tree breeding strategies using best linear unbiased prediction. *Theoretical and Applied Genetics* **96**: 484-493.

LAMBETH, C.C. 1980. Juvenile-mature correlations in Pinaceae and implications for early selection. *Forest Science* **26**: 571-580.

MAGNUSSEN, S. 1988. Minimum age-to-age correlations in early selections. *Forest Science* **34**: 928-938.

MAGNUSSEN, S. 1993. A continuous-time Markov chain for early selection. *Forest Science* **39**: 845-850.

MAGNUSSEN, S. and YANCHUK, A.D. 1993. Selection age and risk: finding the compromise. *Silvae Genetica* **42**: 25-40.

NEWMAN, D.H. and WILLIAMS, C.G. 1991. The incorporation of risk in optimal selection age determination. *Forest Science* **37**: 1350-1364.

SCHNEEBERGER, M., FREEMAN, A.E. and BOEHLJE, M.D. 1982. Application of portfolio theory to dairy sire selection. *Journal of Dairy Science* **65**: 404-409.

SEARLE, S.R. 1965. The value of indirect selection: I. Mass selection. *Biometrics* **21**: 682-707.

SEARLE, S.R. 1982. Matrix algebra useful for statistics. John Wiley & Sons, New York.

VILLANUEVA, B., WRAY, N.R. and THOMPSON, R. 1993. Prediction of asymptotic rates of response from selection on multiple traits using univariate and multivariate best linear unbiased predictors. *Animal Production* **57**: 1-13.

WADA, Y. and KASHIWAGI, N. 1990. Selecting statistical models with information statistics. *Journal of Dairy Science* **73**: 3575-3582.

WEI, M., CABALLERO, A. and HILL, W.G. 1996. Selection response in finite populations. *Genetics* **144**: 1961-1974.

WEI, X. and BORRALHO, N.M.G. 1996. A simple model to describe age trends in heritability in short rotation tree species. P 178-181 *in* DIETERS, M.J., MATHESON, A.C., NIKLES, D.G., HARWOOD, C.E., and WALKER, S.M. (Ed.) Proceedings of Tree Improvement for Sustainable Tropical Forestry, 27 October-1 November 1996, Caloundra, Queensland, Australia.

WHITE, T.L. 1987. A conceptual framework for tree improvement programs. *New Forests* **1**: 325-342.

WHITE, T.L., HODGE, G.R. and DELORENZO, M.A. 1987. Best linear prediction of breeding values in forest tree improvement. P 99-122 *in* Proceedings of the 1986 Workshop on Statistical Considerations in Genetic Testing. University of Florida, Gainesville, Florida.

WHITE, T.L. and HODGE, G.R. 1988. Best linear prediction of breeding values in a forest tree improvement program. *Theoretical and Applied Genetics* **76**: 719-727.

WOOLLIAMS, J.A. and MEUWISSEN, T.H.E. 1993. Decision rules and variance of response in breeding schemes. *Animal Production* **56**: 179-186.

ZOBEL, B.J. and TALBERT, J.T. 1984. Applied forest tree improvement. John Wiley & Sons, New York.

# CHAPTER SEVEN

# GENERAL DISCUSSION

This thesis concentrates on problems faced by tree breeding that can be approached, and benefit, from a multivariate perspective. The first problem is the definition of a breeding objective under three silvicultural regimes for a vertically integrated firm, achieved by developing a model comprising mathematical descriptions of a forest, a sawmill and a pulp mill. The second is the simultaneous estimation of genetic parameters for a variable that is measured in all trees of a progeny test (e.g. growth), while other variables (especially wood properties) are subsampled for cost/efficiency reasons. The third is the analysis of longitudinal data — repeated assessments of a trait at different times — using a variety of statistical models to consider the relationships that result from the sequential nature of the assessments. The final problem is the integration of longitudinal data and family information, assuming different 'true' covariance structures, to determine optimum evaluation time(s) to maximise expected genetic gain per year.

In keeping with the association among chapters the discussion is divided in three sections: breeding objectives (Chapter two), sampling progeny tests (Chapter three), and the exploitation of longitudinal data for early selection (Chapters four, five and six).

## Breeding objectives

The breeding objectives discussed in Chapter two, which include harvest age volume ($\Delta$vol in $m^3$/ha) and average wood density ($\Delta$den in kg/$m^3$), result from a simplified representation of a generic vertically integrated firm (including only forest, sawmill and pulp mill components). The implementation of a breeding objective for the Chilean breeding effort certainly is a more complex situation, which may require the inclusion of other subsystems into the firm (e.g. fibreboards, paper). Nevertheless, it is probable that adding further detail to the model may only increase precision of the estimates, without greatly changing the relative economic values.

Under the silvicultural regimes modelled in this study, which represent different site qualities and cost/income structures, the breeding objectives were 1 $\Delta$vol + 2.38 $\Delta$den for pulp, 1 $\Delta$vol + 1.14 $\Delta$den for intermediate, and 1 $\Delta$vol + 1.23 $\Delta$den for the intensive silvicultural regime. However, the use of a single breeding population with an average objective (1 $\Delta$vol + 1.47 $\Delta$den) may be the best option to maximise profit. The

literature for multiple breeding objectives is not abundant (e.g. DEL BOSQUE GONZÁLEZ and KINGHORN 1990, HOWARTH et al. 1997) and extensions to this study can prove useful to define the most appropriate combination of national breeding strategy (across firms) and the particular requirements for specific firms. MESZAROS et al. (1997) propose the use of genetic algorithms to assist with the design of breeding objectives and strategies for a national industry context.

MEUWISSEN (1998) points out the possibility of including risk — including variance of response, inbreeding and changing market conditions — as a cost in the definition of breeding objectives. Part of the risk is often explored through sensitivity analysis, evaluating the effects of changes to prices using either a distribution of values or a set of scenarios (GROEN 1990, FOGARTY and GILMOUR 1993, GREAVES 1999). Most difficult to predict is the economic effect of new products or technological changes. For example, the Framework Convention of Climate Change acknowledges forests as one of the main greenhouse gas sinks (FCCC 1998). GREAVES (1999) suggests that the introduction of carbon emission taxes might affect the economics of forestry, and therefore the breeding objectives. Carbon sequestration might be considered in the future as a potential source of income when defining forest breeding objectives.

Chapter two does not specifically cover the definition of selection criteria, and for the sake of simplicity assumes they are identical to the breeding objective traits ($\Delta$vol and $\Delta$den). Nevertheless, total tree height (m), diameter at breast height (cm) and wood density — estimated using increment cores or pilodyn, see GREAVES et al. 1996 for example — at early ages are likely selection criteria. Some issues that deserve future attention are: the lack of information on the behaviour of genetic correlations between selection criteria over time, the relationship between performance on progeny tests (several families growing in small plots) and stands (often blocks of individuals from one family or clone), and the effect of highly skewed distributions at rotation age — diameter (or basal area) is often approximated with a Weibull distribution — when predicting future performance based on early assessments.

## Sampling progeny tests

The lack of reliable genetic parameters estimates for traits that describe wood properties is currently a big deficiency in tree breeding, for an example see NYAKUENGAMA et al. (1999) where the heritabilities of four out of eight variables are outside the parameter space. Sampling schemes have a large influence on the quality of genetic parameter estimates. Chapter three shows that in random subsampling the largest biases in parameters are associated with estimates of the genetic correlation between the traits, followed by estimates of the heritability of the subsampled trait. Additionally accuracy of prediction is directly related to the sampling percentage and the heritability of the trait subject to subsampling. Chapter three also shows that performance of truncation subsampling — all families assessed and ranked by growth and a percentage of the top families assessed for wood properties, a common practice in tree breeding programmes — provides very poor results, even when utilising measurements on a greater number of trees than with random subsampling. The analyses reported in Chapter three cover only half-sib families, but it is expected that the results relating to the efficiency of low subsampling intensities will be even more pronounced for families with higher coefficient of relationship (e.g. full-sibs and clones).

While random subsampling of progeny tests has been shown to be appropriate with as few as 9 to 15 trees per family (Chapter three), those numbers can still be too large given the economics of wood processing analysis. Possible options are to develop cheap assessment methods (a successful example is measurement of spiral grain with a custom made tool, SORENSSON et al. 1997) or to use other approaches like parent-progeny relationship, and the use of clonal material. Elliptical sampling (CAMERON and THOMPSON 1986) — sampling the top and bottom extremes of the distribution — combined with parent-progeny regression may provide a good compromise of accuracy versus costs of analysis.

## Longitudinal data and early selection

The central issue for longitudinal data as developed in Chapters four and five is the expression of the additive genetic covariance matrix ($\mathbf{G}_0$) as a correlation matrix ($\mathbf{C}$) pre- and post-multiplied by a diagonal matrix ($\mathbf{S}$) of additive genetic standard deviations, and then modelling the correlation matrix. This representation allows for

heterogeneous variances for different ages, which are frequent during the long progeny testing periods used in tree breeding. These chapters present the widest selection of models reported for genetic data, including unstructured (US), banded correlations (BC), autoregressive (AR), repeatability (RE), random regressions (RR) and uncorrelated (UC) models. All the models are characterised by a reduction in the number of covariance parameters fitted to the data in comparison to US, in order to unveil the data structures created by the repeated assessments. This approach facilitates the analysis of longitudinal data while reducing the risk of non positive definite matrices. The preference is for models with clear biological meaning, e.g. second order polynomial for RR (Chapter four) and AR with age on a natural logarithm scale (Chapter five). The main question posed by the AR model in Chapter five is whether the autocorrelation coefficient is really so high (0.94). This issue points to the need to perform similar analysis on larger data sets.

A problem that became apparent in the analyses concerns model selection techniques and the availability of data sets appropriate for fitting the models. Several methods for model selection have been put forward, mostly variations of a penalised likelihood approach, i.e. a weighted function of log-likelihood value and number of independent parameters fitted in the model, e.g. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (JONES 1993). However, as shown in Chapter five, a small difference in AIC can be associated with large differences to the covariance matrix. Graphical methods (e.g. plotting covariances estimated using the US method versus the ones estimated by alternative methods) should be explored. While highly subjective, these methods can provide an illustration of the effects of parsimonious models. Tree breeding data sets currently may be computationally too demanding for data augmentation techniques to compare covariance matrices (e.g. bootstrap, GOODNIGHT and SCHWARTZ 1997).

Although currently RR models are popular in animal breeding (e.g. JAMROZIK et al. 1997, VAN DER WERF et al. 1998), the results obtained in this thesis are irregular. Working with a small number of assessments (four in Chapter four) RR provide a fitting very close to the US model while using four fewer covariance parameters, and these also provided a function to model the additive genetic covariance matrix. On the contrary, when using a RR model with 10 assessments (Chapter five) the result was a

very poor fitting, although the AIC value was close to that for the US model. Additional work is required to determine the conditions when RR models are a good option for longitudinal data, as well as testing their ability to model other random effects present in the model (e.g. block, plot and residuals). Equally important will be the extension of these models to include data from more than one site and/or more than one trait in the analysis (e.g. ROCHON 1996), situations that are commonplace in operational breeding programs.

Tree breeding strategies are conditioned by the long time interval between plantation establishment and harvesting. In principle, selection of trees for subsequent breeding generations and multiplication (either sexual or vegetative) is more accurate when performed close to rotation age. However, waiting until close to rotation age will dramatically extend generation intervals, and therefore reduce expected gain per year, which is one of the criteria to assess the efficiency of a breeding program. For this reason, tree breeding has favoured the use of early selection, often settling on a conservative one-third or larger fraction of rotation time (FRANKLIN 1979, MCKEAND 1988, MAGNUSSEN and YANCHUK 1993). Further optimisations to selection age rely on better knowledge of the genetics of development, and the use of early indicators of future performance such as genetic markers (CARSON et al. 1996, KERR and GODDARD 1997, XIE and XU 1998). Chapters four to six focus on the former.

BURDON (1989) suggested a likely role for longitudinal data in index selection in tree breeding. Chapter six establishes that using information from collateral relatives provides more gain than including additional assessments on individual trees. Using family information increases predicted gain by 16% to 41%, while multiple assessments increase gain by no more than 7%. Considering longer breeding delays (from 5 to 8 years) slightly reduces expected response to selection per year. These results will depend on the covariance structure, with higher predicted gains for the AR model with age in a natural logarithm scale. While the expected genetic gains from using alternative covariance structures are different, it is not possible to compare them without data from an additional generation of trees. There are two approaches for simulating expected gain for multiple generations: through 'approximated BLUP' equations (VILLANUEVA et al. 1993, KERR 1998) or simulating individual trees (e.g. MULLIN and PARK 1995). Simple equations may include the reduction of additive genetic variance

due to selection (Bulmer effect, BULMER 1971) and have low computational requirements; however, including other genetic considerations (e.g. inbreeding depression) greatly complicates the equations (WEI et al. 1996). Simulation of individual trees allows the straightforward implementation of the Bulmer effect, inbreeding depression, common environmental variation, and mutational variation, but greatly increases computational requirements.

Early selection is part of a larger problem, which is selecting individuals considering risk (variance of response and inbreeding) in a multiple generation context. Concerning variance of response SCHNEEBERGER et al. (1982), SMITH and HAMMOND (1986), and NASH and ROGERS (1996) propose the use of portfolio theory to select individuals with varying levels of accuracy through the use of quadratic programming. MEUWISSEN (1991) use utility theory to balance expected response and its variance, while WOOLLIAMS and MEUWISSEN (1993) develop Bayesian selection rules considering the estimated breeding values and the predicted error variance matrix. Similarly, for the problem of constraining inbreeding WRAY and GODDARD (1994) and BRISBANE and GIBSON (1995) suggest selecting individuals considering their average parental breeding value, with a penalty (or cost) for their average genetic relationship, thus reducing the level of inbreeding due to selection. In an alternative method VILLANUEVA and WOOLLIAMS (1997) propose a method for maximising genetic gain with constraints on inbreeding using index selection, and GRUNDY et al. (1998) develop an alternative algorithm for maximising genetic response while constraining the rate of inbreeding. All these approaches consider only one round of selection at the time. Selecting across multiple generations is a combinatorics problem, with no closed solution unless enumerating all possible combinations. MEUWISSEN and WOOLLIAMS (1994) propose the use of simulated annealing, while HAYES (1997), KINGHORN (1999) and SHEPHERD and KINGHORN (1999) put forward the use of evolutionary algorithms (used in the context of mate allocation). None of the methods — but complete enumeration — guarantees a general optimum, but many may identify solutions close to the optimum.

## General Conclusions

The main conclusions from this thesis are:

- The breeding objectives for three silvicultural regimes were similar, and a single breeding objective appears to provide more expected gain than the use of specialist objectives.
- Subsampling progeny tests is subject to the Law of Diminishing returns, and more than 15 trees per family of size 30 do not provide large gains in accuracy of genetic parameters and prediction of expected gain.
- The use of covariance structures reduces the risk of non-positive definite additive genetic matrices, while reducing computational demands for the analyses and providing a description of the genetic control of a trait over time.
- Model selection techniques based on penalised log-likelihood are not completely appropriate to discriminate between competing models with different covariance structures.
- Including family information in the prediction of breeding values provides higher expected response to selection than integrating repeated measurements in the prediction procedure. A combination of family information and repeated assessments is likely to provide the highest genetic gains.

## Literature cited


BULMER, M.G. 1971. The effect of selection on genetic variability. *The American Naturalist* **105**: 201-211.

BURDON R.D. 1989. Early selection in tree breeding: principles for applying index selection and inferring input parameters. *Canadian Journal of Forest Research* **19**: 499-504.

CAMERON, N.D., and THOMPSON, R. 1986. Design of multivariate selection experiments to estimate genetic parameters. *Theoretical and Applied Genetics* **72**: 466-476.

CARSON, M.J., CARSON, S.D., RICHARDSON, T.E., WALTER, C., WILCOX, P.L., BURDON, R.D., and GARDNER, R.C. Molecular biology applications to forest trees — fact, or fiction? P 272-281 *in* DIETERS, M.J., MATHESON, A.C., NIKLES, D.G., HARWOOD, C.E., and WALKER, S.M. (Ed.) "Tree Improvement for Sustainable Tropical Forestry". Proceedings of the QFRI-IUFRO Conference, 27 October- 1 November, Caloundra, Queensland, Australia.

DEL BOSQUE GONZÁLEZ, A.S., and KINGHORN, B.P. 1990. Implications of different selection objectives within open nucleus breeding schemes. P 95-102 *in* Australian Association of Animal Breeding and Genetics. Proceedings of the Eighth Conference, 5-9 February, Hamilton and Palmerston North, New Zealand.

FCCC. 1998. Report of the conference of the parties on its third session, held at Kyoto from 1 to 11 December 1997. Part two: action taken by the conference of the parties at its third session.

FOGARTY, N.M., and GILMOUR, A.R. 1993. Sensitivity of breeding objectives to prices and genetic parameters in Australian Corriedale and Polwarth dual-purpose sheep. *Australian Journal of Experimental Agriculture* **33**: 259-268.

FRANKLIN, E.C. 1979. Model relating levels of genetic variance to stand development of four North American conifers. *Silvae Genetica* **28**: 207-212.

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

GREAVES, B.L. 1999. Estimating an economic breeding objective for radiata pine grown for structural sawn-timber and liner-board. *Canadian Journal of Forest Research* (submitted).

GREAVES, B.L., BORRALHO, N.M.G., RAYMOND, C.A., and FARRINGTON, A. 1996. Use of a Pilodyn for the indirect selection of basic density in *Eucalyptus nitens*. *Canadian Journal of Forest Research* **26**: 1643-1650.

GROEN, A.F. 1990. Influences of economic circumstances on the economic revenue of cattle breeding programmes. *Animal Production* **51**: 469-480.

GRUNDY, B., VILLANUEVA, B., and WOOLLIAMS, J.A. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. P 355-358, Vol. 25 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production. 11-16 January, Armidale, New South Wales, Australia.

HAYES, B.J., SHEPHERD, R.K., and NEWMAN, S. 1997. Selecting mating pairs with genetic algorithms. P 108-112 *in* Part One, Association for the Advancement of Animal Breeding and Genetics. Proceedings of the Twelfth Conference, 6-10 April, Dubbo, New South Wales, Australia.

HOWARTH, J.M., GODDARD, M.E., and KINGHORN, B.P. 1997. Breeding strategies for targeting different breeding objectives. P 99-102 *in* Part One, Association for the Advancement of Animal Breeding and Genetics. Proceedings of the Twelfth Conference, 6-10 April, Dubbo, New South Wales, Australia.

JAMROZIK, J., KISTEMAKER, G.J., DEKKERS, J.C.M., and SCHAEFFER, L.R. 1997. Comparison of possible covariates for use in random regression models for analyses of test days yields. *Journal of Dairy Science* **80**: 2550-2556.

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

KERR, R.J. 1998. Asymptotic rates of response from forest tree breeding strategies using best linear unbiased prediction. *Theoretical and Applied Genetics* **96**: 484-493.

KERR, R.J., and GODDARD, M.E. 1997. Comparison between the use of MAS and clonal tests in tree breeding programmes. P 297-303 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

KINGHORN, B.P., and SHEPHERD, R.K. 1999. Mate selection for the tactical implementation of breeding programs. P 130-133 *in* Proceedings of the 13th Conference of the Association for the Advancement of Animal Breeding and Genetics. 4-7 July, Australia.

MCKEAND, S.E. 1988. Optimum age for family selection for growth in genetic tests of loblolly pine. *Forest Science* **34**: 400-411.

MAGNUSSEN, S., and YANCHUK, A.D. 1993. Selection age and risk: finding the compromise. *Silvae Genetica* **42**: 25-40.

MESZAROS, S.A., BANKS, R.G., KINGHORN, B.P., and SHAFTO, A.M. 1997. Design considerations in development of breeding strategies in a complex national industry context. P 95-98 *in* Part One, Association for the Advancement of Animal Breeding and Genetics. Proceedings of the Twelfth Conference, 6-10 April, Dubbo, New South Wales, Australia.

MEUWISSEN, T.H.E. 1991. Expectation and variance of genetic gain in open and closed nucleus and progeny testing schemes. *Animal Production* **53**: 133-141.

MEUWISSEN, T.H.E. 1998. Risk management and the definition of breeding objectives. P 347-354, Vol. 25 *in* Proceedings of the 6th World Congress of Genetics Applied to Livestock Production. 11-16 January, Armidale, New South Wales, Australia.

MEUWISSEN, T.H.E., and WOOLLIAMS, J.A. 1994. Maximizing genetic response in breeding schemes of dairy cattle with constraints on variance of response. *Journal of Dairy Science* 77: 1905-1916.

MULLIN, T.J., and PARK, Y.S. 1995. Stochastic simulation of population management strategies for tree breeding: a new decision-support tool for personal computers. *Silvae Genetica* **44**: 132-141.

NASH, D.L., and ROGERS, G.W. 1996. Risk management in herd sire portfolio selection: a comparison of rounded quadratic and separable convex programming. *Journal of Dairy Science* **79**: 301-309.

NYAKUENGAMA, J.G., EVANS, R., MATHESON, C., SPENCER, D., and VINDEN, P. 1999. Wood quality and quantitative genetics of *Pinus radiata* D. Don: fibre traits and wood density. *Appita Journal* **52**: 348-350.

ROCHON J. 1996. Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics* **52**: 740-750.

SCHNEEBERGER, M., FREEMAN, A.E., and BOEHLJE, M.D. 1982. Application of portfolio theory to dairy sire selection. *Journal of Dairy Science* **65**: 404-409.

SHEPHERD, R.K., and KINGHORN, B.P. 1999. Algorithms for mate selection. P 126-129 *in* Proceedings of the 13th Conference of the Association for the Advancement of Animal Breeding and Genetics. 4-7 July, Australia.

SMITH, S.P., and HAMMOND, K. 1986. The ramifications of portfolio and utility theories for dairy breeding programs. P 101-105, Vol. 9 *in* Proceedings of the 3th World Congress of Genetics Applied to Livestock Production. 16-22 July, Lincoln, Nebraska, USA.

SORENSSON, C.T., BURDON, R.D., COWN, D.J., JEFFERSON, P.A., and SHELBOURNE, C.J.A. 1997. Incorporating spiral grain into New Zealand's radiata pine breeding programme. P 180-191 *in* BURDON, R.D., and MOORE, J.M. (Ed.) "IUFRO '97 Genetics of Radiata Pine". Proceedings of NZFRI-IUFRO Conference 1-4 December and Workshop 5 December, Rotorua, New Zealand. FRI Bulletin 203.

VAN DER WERF, J.H.J., GODDARD, M.E., and MEYER, K. 1998. The use of covariance functions and random regression for genetic evaluation of milk production based on test day records. *Journal of Dairy Science* **81**: 3300-3308.

VILLANUEVA, B., and WOOLLIAMS, J.A. 1997. Optimization of breeding programmes under index selection and constrained inbreeding. *Genetical Research* **69**: 145-158.

VILLANUEVA, B., WRAY, N.R., and THOMPSON, R. 1993. Prediction of asymptotic rates of response from selection on multiple traits using univariate and multivariate best linear unbiased predictors. *Animal Production* **57**: 1-13.

WEI, M., CABALLERO, A., and HILL, W.G. 1996. Selection response in finite populations. *Genetics* **144**: 1961-1974.

WOOLLIAMS, J.A., and MEUWISSEN, T.H.E. 1991. Decision rules and variance of response in breeding schemes. *Animal Production* **56**: 179-186.

WRAY N.R., and GODDARD, M.E. 1994. Increasing long-term response to selection. *Genetics, Selection, Evolution* **26**: 431-451.

XIE, C. and XU, S. 1998. Efficiency of multistage marker assisted selection in the improvement of multiple quantitative traits. *Heredity* **80**: 489-498.

## Curriculum Vitae

Luis Alejandro Apiolaza was born on April 5th 1967 in Concepción, Chile. He attended primary and secondary schools in Chile, Venezuela and Argentina. In 1987 Luis started tertiary education at the School of Forest Sciences, Universidad de Chile, Santiago. He completed his Bachelor of Forestry Science with first class honours in 1992, and obtained his professional title in 1994, with the thesis "Genetic evaluation of juvenile growth of *Eucalyptus camaldulensis* Dehnh. at Mel-Mel and Longotoma, V Región". Luis was awarded the 1994 "School of Forest Sciences Award" for the best student considering academic performance and quality of the thesis. Between April 1993 and February 1996 he worked as Research Officer in Quantitative Genetics in the Chilean Tree Improvement Co-operative, Valdivia. He spent August 1994 in the School of Forest Resources and Conservation, University of Florida, USA. In 1996 he was awarded a New Zealand Official Development Assistance scholarship to pursue a PhD degree at Massey University. Upon completion of the PhD Luis will undertake a three-year post-doctoral position with the Co-operative Research Centre for Sustainable Forestry at the School of Plant Science, University of Tasmania, Australia.