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] **& 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
[*]Corresponding author

Received January 7, 1998; accepted February 26, 1999

ABSTRACT

Bivariate half-sib family data 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 comprising 200 unrelated families with 30 individuals each, 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 breeding values, 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 breeding values 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.

Keywords: subsampling, half-sibs, genetic parameters, breeding values, response to selection.

INTRODUCTION

Tree performance is typicalily a multi-trait function. For efficient genetic improvement good 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 good genetic parameters estimates that are often needed for constructing reliable selection indices, can allow good 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 & 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 the 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 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 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 breeding values for backwards (or parental) selection. Therefore, for each progeny test was considered a family ('sire'[1]) model:

$$y = X m + Z f + e$$

where $y = (y_1\grave{}\ y_2\grave{})$ represents the vector of phenotypic observations for traits 1 and 2, $\boldsymbol{X = X_1 \oplus X_2}$ and $\boldsymbol{Z = Z_1 \oplus Z_2}$ are known incidence matrices for fixed and random effects respectively, $\boldsymbol{m = (m_1\grave{}\ m_2\grave{})\grave{}}$ and $\boldsymbol{f = (f_1\grave{}\ f_2\grave{})\grave{}}$ are vectors of unknown trait means and random family effects respectively, $\boldsymbol{e}$ is the vector of random residuals, ` is the transpose operation, and ⊕ is the direct sum operation. The expected value (*E*) and variance (*V*) of the model equation terms are:

$$E\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{e} \\ \boldsymbol{f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Xm} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \quad \text{and} \quad V\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{e} \\ \boldsymbol{f} \end{bmatrix} = \begin{bmatrix} \tfrac{1}{4}\boldsymbol{ZG\grave{Z}}+\boldsymbol{R} & \boldsymbol{R} & \tfrac{1}{4}\boldsymbol{ZG} \\ \boldsymbol{R} & \boldsymbol{R} & \boldsymbol{0} \\ \tfrac{1}{4}\boldsymbol{G\grave{Z}} & \boldsymbol{0} & \tfrac{1}{4}\boldsymbol{G} \end{bmatrix}$$

where:

$\boldsymbol{G = G_0 \otimes I_N}$ is the additive genetic variance-covariance matrix,

$$\boldsymbol{G_0} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} \end{bmatrix} = 4\begin{bmatrix} \sigma^2_{f_1} & \sigma_{f_1f_2} \\ \sigma_{f_1f_2} & \sigma^2_{f_2} \end{bmatrix}$$

where ⊗ denotes the direct (Kronecker) product operation, and $\boldsymbol{I_N}$ 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 $r_g h_1 h_2$.

$\boldsymbol{R}$ 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, $\boldsymbol{R}$ cannot be expressed as a direct product. For individuals with records for both traits $\boldsymbol{R_0}$ diag $\{\sigma^2_{e_1}, \sigma^2_{e_2}\}$; while for individuals with records only for trait 1, the matrix $\boldsymbol{R_0}$ collapses to the scalar $\sigma^2_{e_1}$.

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 sampled for trait 2.

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

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

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

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

with $\sigma_{f_i}^2$ and $\sigma_{f_i f_j}$ as the among-families estimated variance and covariance components respectively.

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):

$$\frac{g_{\hat{p}_i}}{(\sigma_{\hat{p}_i})^{\frac{3}{2}}} > \text{crit}_{\alpha,\,n}$$

where $\sigma_{\hat{p}_i}$ is the population standard deviation of the n values of p_i for a given combination of parameters, and $\text{crit}_{\alpha,n}$ is the tabulated critical value for a nominal probability of a comparisonwise type-I error, α, and 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 α and n – 1 degrees of freedom.

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

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

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

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

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

The breeding objective (H) comprised a linear function of the additive genetic value for traits 1 and 2 (f_1 and f_2). Selection was performed using an index (I), which included the bivariate-predicted breeding values 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 (ΔG_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):

$$\Delta_c G_i = i\grave{w}\boldsymbol{T}_i(\grave{w}\boldsymbol{S}w)^{-1/2}$$

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

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

RESULTS AND DISCUSSION

Random Subsampling

The results are presented separately for estimation of genetic parameters, prediction of breeding values, 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 assume relative economic weights 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.

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 breeding values

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 breeding values was assessed using the correlation between breeding values 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 (Fig. 1, Table 2).

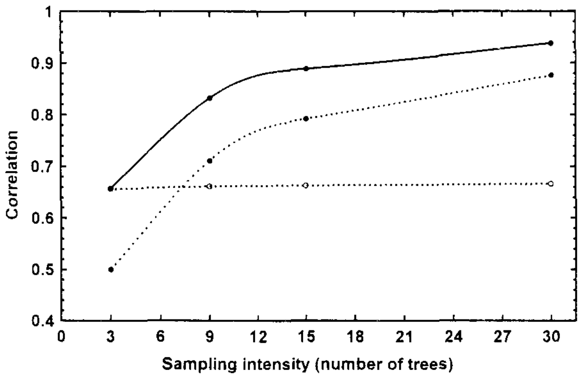

Figure 1. Correlation between breeding values predicted using 3, 9, 15 and 30 observations and those simulated, considering $r_{\hat{f}_i f_{si}}$ for $h_1^2 = 0.1$, $h_2^2 = 0.4$ and $r_g = -0.3$ (···o···), and $r_{\hat{f}_i f_{si}}$ for $h_1^2 = 0.1$, $r_g = -0.3$ and $h_2^2 = 0.4$ (···•···) or $h_2^2 = 0.8$ (–•–).

Table 1. Bias, coefficient of variation (CV) and skewness (Skew) 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	0.001	0.020	0.332	–0.084	0.019	–0.526	0.640	0.117	–0.398

Tr = special case with truncation subsampling, using non-random selection of the assessed families.

Table 2. Correlation between breeding values 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

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 (Fig. 1). Simultaneously, a higher simulated correlation r_g between the traits (either positive or negative) gave higher correlations with predicted breeding values (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).

Even when 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 & MEYER 1986). However, the effect of subsampling on trait 2 was small compared with the results for $r_{\hat{f}_1 f_{s2}}$ (Fig. 1, Table 2). The effect of including trait 2 depended on the value of r_g; thus the highest $r_{\hat{f}_2 f_{s1}}$ 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, $r_{\hat{f}_2 f_{s1}}$ ranged from 0.65 to 0.85, indicating a considerable agreement between the selection of parents under the different subsampling intensities.

Response to selection

To study the effect of subsampling, both types of response 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 (ΔG_H) and correlated ($\Delta_c G_i$) re-

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

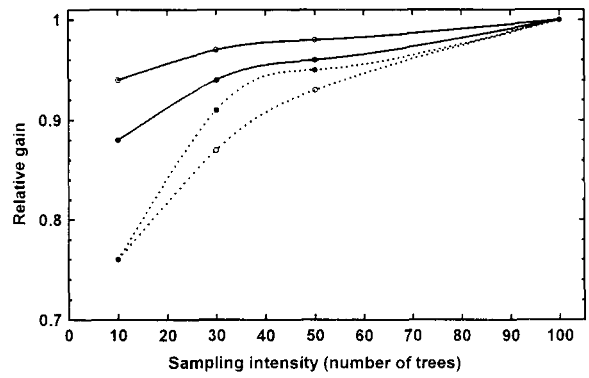

Figure 2. Average relative gain for different subsampling intensities, expressed as the ratio of response to the predicted direct response using the true parameters. Results presented for relative economic weights 2 (trait 1) and 1 (trait 2), considering $h_1^2 = 0.1$ and $h_2^2 = 0.4$ (···o···), $h_1^2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

sponses. Low subsampling intensities consistently depressed the predicted response compared to the case with true genetic parameters (Fig. 2). Direct response was very dependent on the heritabilities of the traits (Fig. 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 (Fig. 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 (Fig. 2). With further subsampling the difference converges to 0, with diminishing cost-efficiency.

Correlated responses showed more dramatic changes when trait 2 was subject to subsampling, especially at the lowest intensity (Fig. 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, Fig. 3) and the effect of genetic correlation was relevant only for low subsampling intensities.

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 subsampling 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).

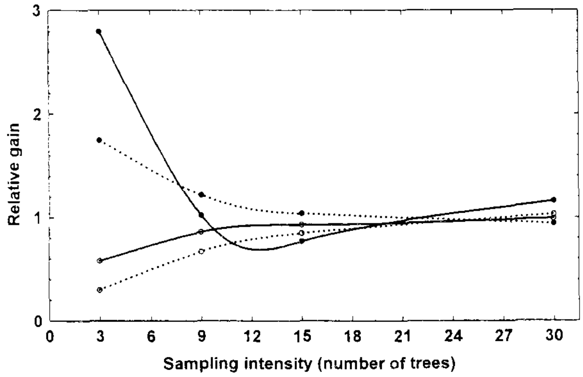

Figure 3. Correlated relative responses to selection, expressed as the ratio of response to the predicted correlated response, using the true parameters and relative economic weights 2 (trait 1) and 1 (trait 2). Correlated responses for trait 1 (···•···) and 2 (···o···) for $h_1^2 = 0.1$, $h_2^2 = 0.4$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

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	0.95	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 percent age 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 h_2^2 and r_g (Table 1). An extreme case is that of r_g, where the estimated parameter averaged 0.34 rather than –0.3. When predicting breeding values 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 breeding values 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 the analysis tools, REML is becoming the preferred analysis method in forest genetics, mainly because 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 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 the prediction of breeding values for 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, available family size, 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 breeding value and response to selection.

ACKNOWLEDGMENTS

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.

REFERENCES

BORRALHO, N. M. G., COTTERILL, P. P. & 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. & 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. & 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. & BORRALHO, N. M. G. 1996: The influence o f 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. & 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, 462 pp.

HUBER, D. A., WHITE, T. L. & 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. & 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, 230 pp.

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

PATTERSON, H. D. & 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. & 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. & SORENSSON, C. T. 1997: Developing breeding objectives for radiata pine in New Zealand. *In:* Proceedings of the IUFRO Conference S2.02.19, Genetics of radiata pine. 1–4 Dec, 1997. Rotorua, New Zealand. p. 160–168.

SNEDECOR, G. W. & COCHRAN, W. G. 1980: Statistical Methods, 7th edition. The Iowa State University Press, Iowa, 505 pp.

THOMPSON, R. & 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, 481 pp.

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. & 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 additive genetic value of a family for trait i (in this case i = 1,2) has the form:

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

where 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 are:

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

where

$$\boldsymbol{P} = \begin{bmatrix} V(\bar{P}_1) & Cov(\bar{P}_1,\bar{P}_2) \\ Cov(\bar{P}_1,\bar{P}_2) & V(\bar{P}_2) \end{bmatrix}$$

and

$$\boldsymbol{q}_i = \begin{bmatrix} Cov(\bar{P}_1 f_i) \\ Cov(\bar{P}_2 f_i) \end{bmatrix}$$

moreover

$$V(\bar{P}_1) = \frac{\sigma^2_{P_1}}{n_1}[1 + (n_1 - 1)¼h_1^2]$$

$$V(\bar{P}_2) = \frac{\sigma^2_{P_2}}{n_2}[1 + (n_2 - 1)¼h_2^2]$$

$$Cov(\bar{P}_1 f_1) = \frac{\sigma_{a_1a_i}}{2}$$

$$Cov(\bar{P}_2 f_1) = \frac{\sigma_{a_2a_i}}{2}$$

and

$$Cov(\bar{P}_1,\bar{P}_2) = \frac{\sigma_{P_1P_2}}{n_1} + \frac{(n_2 - 1)¼\sigma_{a_1a_2}}{n_1}$$

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

The bivariate predicted breeding values for traits 1 and 2 (f_1 and f_2) are combined into an index I. This index is used as a prediction of the breeding objective H , which includes the additive genetic values (f_l 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):

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

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

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

$$\boldsymbol{S} = \begin{bmatrix} \hat{c}_1\boldsymbol{P}\hat{c}_1 & \hat{c}_1\boldsymbol{P}\hat{c}_2 \\ \hat{c}_1\boldsymbol{P}\hat{c}_2 & \hat{c}_2\boldsymbol{P}\hat{c}_2 \end{bmatrix}$$

and

$$\boldsymbol{T}_i = \begin{bmatrix} \hat{c}_1\boldsymbol{q}_i \\ \hat{c}_2\boldsymbol{q}_i \end{bmatrix}$$

correlated response of trait i reduces to:

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

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

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