Abstract

BIN XIANG Genetic Analysis of Diallel Tests of Loblolly Pine (*Pinus taeda* L.) (Under the direction of Dr. Bailian Li)

A new approach was developed for analyzing diallel tests with SAS PROC MIXED and PROC IML. The new method can estimate variance components, obtain BLUE (best linear unbiased estimators) of fixed effects and BLUP (best linear unbiased predictors) of random genetic effects simultaneously. A new formula based on BLUP was further developed to predict individual tree breeding values. This new analytical method was validated using computer simulation and was compared with other existing programs.

To analyze disconnected diallel mating designs with more than one diallel, simulated data generated with known parameters were analyzed using BLUP to compare three alternative models, which include diallel as fixed effect (Model 1), random effect (Model 2) or no diallel effect (Model 3). Both Model 1 and Model 3 produced unbiased GCA (general combining ability) variance estimates, while Model 2 resulted in downward biased GCA variance estimate. The accuracy of BLUP prediction for three models was very close, with Model 3 slightly better than the other two.

Statistical approaches were also evaluated for combining multiple disconnected diallel test series in a given region. The best GCA sample variance prediction in the class of linear combination of local variance estimates was derived. Simulation study showed that a checklot adjustment was very critical to improve the prediction of genetic values obtained using BLUP analysis. Additional adjustment with improved GCA sample variance prediction could improve the correlation slightly beyond checklot adjustment.

Analysis of annual measurement through age 8 from a total of 275 parents, 690 full-sib families from 23 diallel tests of loblolly pine in Northern, Coastal and Piedmont test regions showed: 1) dominance variance was small (20-40% of total genetic variance) relative to additive variance; 2) heritability increased over time, and the magnitude of heritabilities for diameter at breast height (DBH) and volume was comparable with the corresponding heritabilities for height; 3) DBH and volume had higher genetic correlation with 8-year volume than height.

Genetic gain prediction in year-8 volume for selection on height and volume indicated that: 1) selection on volume yielded more gain than selection on height; 2) Coastal population had the greatest correlated response, followed by Piedmont and Northern population; 3) family plus within family selection based on total genetic component can capture the most genetic gain; 4) for all selection methods, additional gain (10-40%) can be achieved by capturing non-additive genetic component.

Selection efficiency study of height and volume for three test regions indicated that earlier selection appeared to be more efficient than direct selection on year-8 volume in most selection methods. Family selection can be performed at age 2 or 3 for height and at age 4 for DBH and volume. Combined selection (family plus within family) was highly efficient at age 3 or 4.

GENETIC ANALYSIS OF DIALLEL TESTS OF LOBLOLLY PINE (Pinus taeda L.)

By

Bin Xiang

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

DEPARTMENT OF FORESTRY

Raleigh

December, 2000

Approved By:

Gary R. Hodge

Eugene J. Eisen

Bailian Li
CHAIR OF ADVISORY COMMITTEE

Roger L Berger

BIOGRAPHY

Bin Xiang was born on December 29, 1970 in Yuyao, Zhejiang Province, China. At the age of seven, he began to receive education at a neighborhood primary school.

The author completed his undergraduate studies at Nanjing University with a Bachelor of Science in Ecology in 1992. At the same year, his graduate studies started at the Institute of Botany, Chinese Academy of Sciences. After three years' study and research in plant physiological ecology, he graduated with a Master of Science in Botany in 1995.

In August 1996, the author enrolled in North Carolina State University to pursue his doctoral studies in Forest Quantitative Genetics. During the studies, he worked as a graduate research assistant for Tree Improvement Program. In May 2000, he obtained a Master's degree in Statistics from North Carolina State University.

ACKNOWLEGEMENTS

The pursuing of PhD degree would not be possible without the guidance, support, coaching and mentoring of my advisor Dr. Bailian Li. He not only guided me through this process but also taught many valuable lessons of the life.

I would like to extend my appreciation to committee members, Dr. Eugene Eisen, Dr. Gary Hodge and Dr. Roger Berger for their time and effort in pursuit of this work. They were always there to offer helping hand throughout the process. I am also grateful to Department of Forestry and the Tree improvement Program at North Carolina State University for funding and supporting this work.

I extend special thanks to my parents Shougen Xiang and Xinhui Wang for their unselfish love and constant emotional support through these years of studies.

Lastly, my sincere thanks go to all my friends here in Raleigh. Their supports gave me the strength to overcome many obstacles.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

General Introduction

Loblolly pine is the most important commercial tree crop in the southern United States. It produces over half of the total southern pine wood volume, and it accounts for about 80 percent of all southern pine seedling production in the United States (Lantz et al. 1987). It will become an increasingly important source for softwood fiber for pulp and timber in the future (Mckeand et al. 1997, Li et al 1999).

The North Carolina State University - Industry Cooperative Tree Improvement Program (NCSU-ICTIP) has completed 45 years of loblolly pine tree improvement in the southeastern United States. Through the first two cycles of breeding, testing and selection substantial genetic gains have been achieved. The cooperative's tree improvement program for loblolly pine is now moving into its third generation (Li *et al.* 1996, Mckeand et al. 1997, Li et al 1999). To further implement the third generation and future breeding plans, it is necessary to thoroughly analyze and evaluate current data accumulated from the first and second-generation progeny tests of NCSU-ICTIP.

The analysis of data from the first generation tests indicated time trends in genetic parameters for tree height and suggested that, if a single measurement is used, measurement at age 6 and selection one year later would maximize the gains per year as well as increase the financial returns from seed orchards (McKeand 1988, Balocchi *et al.* 1993).

Well-balanced data from second-generation genetic testing are now available to estimate genetic parameters at early ages more accurately and precisely. The mating design used in second-generation testing is a diallel design, which is a widely used mating design for forest tree improvement programs. With respect to four main objectives of genetic testing (Zobel and Talbert 1984), it can yield information on general combining ability (GCA) of all parents, specific combining ability (SCA) of all crosses, and variance components and heritability, and provide the progeny population for the next generation of breeding. The expected genetic gain can also be predicted.

With better estimates of parameters, time trends of selection efficiency can be calculated for different selection methods. These selection methods may include mass selection, family selection, within family selection and index selection with combined family and within family information. The optimal ages for different selection methods may be estimated more reliably and may not be the same as those obtain from the first generation tests. More importantly, with new and better information, alternative selection strategies such as volume selection BLUP selection and their optimal selection ages can also be investigated to get maximum genetic gains per unit of time.

Objectives of this thesis research project are to:

1. Evaluate and develop analytical methodologies for genetic parameter estimation and genetic gain prediction based on diallel tests of loblolly pine,

2. Examine time trends of genetic parameters for different traits (height, DBH and volume) to determine selection efficiency and optimal ages for different selection methods, and

3. Evaluate different selection strategies to maximize the genetic gains.

Chapter 1 A New Mixed Analytical Method for Genetic Analysis of Diallel Data

Abstract

Diallel mating is a popular mating design used for tree breeding programs, but its unique feature of a single observation with two levels of the same main effect, general combining ability (GCA), makes it difficult to analyze with any standard statistical programs. A new approach using the SAS PROC MIXED is developed in this study for analyzing genetic data from diallel mating. By first constructing dummy variables for GCA effects with SAS IML, a new method using PROC MIXED was developed to estimate variance components, obtain BLUE (best linear unbiased estimators) of fixed effects and BLUP (best linear unbiased predictors) of random genetic effects (GCA and SCA effects) simultaneously. A new formula is derived based on BLUP methodology to predict individual tree breeding values. With the results provided by PROC MIXED, precise breeding values can easily be calculated in SAS IML for every individual tree. This new analytical method was validated using computer simulated data with known genetic parameters and was compared with other existing programs.

Keyword: Diallel mating, General combining ability (GCA), Specific combining ability (SCA), Best linear unbiased prediction (BLUP).

Introduction

Diallel mating designs, especially half or partial diallels, are widely used mating designs in crop and tree breeding programs (Yanchuk, 1996; Huber, 1992). The diallel mating design is popular because it can yield information on general combining ability (GCA) of all parents, specific combining ability (SCA) of all crosses, and genetic variance components and heritability. It also provides a pedigreed progeny population for advanced selection and breeding program (Zobel and Talbert 1984).

While several analytical methods including ordinary least squares (OLS), general least squares (GLS), best linear prediction (BLP) and best linear unbiased prediction (BLUP) may be used for analysis of diallel tests (Borralho 1995, White and Hodge 1988), practical problems remain in using these statistical tools. The unique feature of half-diallel design, which hinders analysis with many existing statistical packages, is that a single observation contains two levels of the same main effect. The analysis is not simple even in the case of balanced data, and complications increase when there are missing plots or missing crosses (Dean 1988, Huber 1992). Special packages have been developed for some uses, such as DIALL (Schaffer and Usanis, 1969), DAG (Dean 1994) and GAREML (Huber 1994) etc. Though these packages can be used to analyze the diallel data, they usually have unfriendly user interfaces, limited options, and require special format of input data and limited options. Most of them cannot be used to estimate individual breeding value. Other limitations include the limited data size that the package can handle and less flexibility in defining analytical models and choosing options to estimate the variance components. The Statistical Analysis System (SAS) is a powerful computer program for almost all analyses (SAS institute 1986), but so far, analysts have not been able to use it directly to analyze diallel genetic data. The newly added SAS procedure (PROC MIXED) provides the flexibility for mixed model analysis and BLUP prediction (Littell et al 1996), but it still can not handle GCA as a main effect in diallel analysis.

In this paper, a new approach is presented in which dummy variables are created for GCA effects in SAS PROC IML. The PROC MIXED procedure is used to estimate variance components, obtain BLUE (best linear unbiased estimators) of fixed effects and BLUP (best linear unbiased predictors) of random genetic effects (GCA and SCA effects) simultaneously.

If the theoretical basis for this approach is correct, SAS PROC MIXED can be widely used and accepted as a standard analytical tool for diallel genetic analysis. To examine the validity of this new method for diallel analysis, a simulation study was carried out in this paper.

Methods

Theoretical consideration

Linear mixed model

The statistical analysis uses individual tree measurements and follows a common scalar linear model for a half diallel mating design with randomized complete block design at multiple location:

$$Y_{ijklm} = \mu + T_i + B_{j(i)} + G_k + G_l + S_{kl} + TG_{ik} + TG_{il} + TS_{ikl} + P_{ijkl} + E_{ijklm} \quad (1)$$

where,

Y_{ijklm} is the *m*th observation of the *j*th block within *i*th test for the *kl*th cross;

μ is the overall mean;

T_i is the fixed effect of *i*th test, *i*=1 to 5;

$B_{j(i)}$ is the *j*th block within *i*th test, *j*=1 to 4;

G_k or G_l is the GCA effect of the *k*th female or *l*th male (*k*,*l*=1, … , 6; *k*<*l*) ~ NID(0, σ^2_{GCA});

S_{kl} is the SCA effect of *k*th and *l*th parents ~ NID(0, σ^2_{SCA});

TG_{ik} or TG_{il} is the *i*th test by the *k*th female or *l*th male GCA interaction ~ NID(0, $\sigma^2_{TEST*GCA}$);

TS_{ikl} is the *i*th test by *k*th and *l*th parents SCA interaction ~ NID(0, $\sigma^2_{TEST*SCA}$);

P_{ijkl} is the random plot effect for the *kl*th cross in the *j*th block within *i*th test ~ NID(0, σ^2_{PLOT});

E_{ijklm} is the within plot error term ~ NID(0, σ^2_E).

All effects except overall mean and test are considered random and independently distributed. The model is a typical mixed model, which can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \qquad (2)$$

where **Y** is the vector of observations, **X** and **Z** are the known design matrices, $\boldsymbol{\beta}$ is the unknown vector of fixed-effects, $\boldsymbol{\gamma}$ is the unknown vector of random-effects including GCA and SCA effects and $\boldsymbol{\varepsilon}$ is the unobserved vector of random errors. $\boldsymbol{\varepsilon}$ and $\boldsymbol{\gamma}$ are assumed to be normally distributed with

$$\mathbf{E}\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\varepsilon}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} \qquad \mathbf{Var}\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\varepsilon}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$$

The variance of **Y** is therefore **ZGZ’** + **R** (SAS Inst. inc., 1996). In most genetic analyses, we also make additional assumptions for $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ such that **G** is a diagonal matrix and $\mathbf{R}=\sigma^2 I_n$.

However, due to the unique feature of diallel designs, the GCA effects can not be classified, i.e. the corresponding columns in **Z** can not be constructed automatically in any standard SAS procedures (e.g. GLM, VARCOMP, MIXED). Hence diallel data cannot be analyzed directly in any of these procedures, and special software packages (DIALL, GAREML, etc.) are often used. To overcome this problem, dummy variables are constructed for each parent with the SAS PROC IML procedure and then the dummy variables are used for analysis with the PROC MIXED procedure (see details in Appendix 1).

Prediction of individual tree breeding value

An individual breeding value for each observation in model (Equ.1) can be further divided into:

$$A_{ijklm} = G_k + G_l + Aw_{ijklm} \quad (3)$$

For each individual observation, the model is

$$Y_{ijklm} = A_{ijklm} + E^*_{ijklm} = G_k + G_l + Aw_{ijklm} + E^*_{ijklm} \qquad (4)$$

where, A_{ijklm} is the breeding value of the *l*th observation of the *j*th block within *i*th test for the *kl*th cross;

G_k or G_l is the GCA effect of the *k*th female or *l*th male (the same as in model 1) ~ NID(0, σ^2_{GCA});

Aw_{ijklm} is the random additive genetic effect of the *l*th observation of the *j*th block within *i*th test for the *kl*th cross ~ NID(0, σ^2_{AW}).

E^*_{ijklm} contains any effects other than additive genetic affects, including nonadditive genetic affects and random environmental affects etc.

Aw_{ijklm} can be called the within family breeding value, which is the part of breeding value contained in random effects other than G_k and G_l. This genetic model can also be written in matrix form:

$$\mathbf{A} = \mathbf{G} + \mathbf{A_w} \qquad \text{and} \qquad \mathbf{Y} = \mathbf{G} + \mathbf{A_w} + \mathbf{E}^* \qquad (5)$$

where **G** is vector of the sum of GCAs from male and female parents, $\mathbf{A_w}$ is the vector of within family breeding values, and **E*** is the vector of the sum of other effects.

Since G_k, G_l and Aw_{ijklm} are defined as additive genetic effects, they are assumed to be independent of each other and independent of E^*_{ijklm}. So we have

$$\mathrm{Cov}(\mathbf{G}, \mathbf{E}^*) = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{A_w}, \mathbf{G}) = \mathbf{0} \qquad \text{and } \mathrm{Cov}(\mathbf{A_w}, \mathbf{E}^*) = \mathbf{0}$$

The variance of A_{ijklm} is the total additive genetic variance σ^2_A (in Equ. 4) and

$$\sigma^2_{GCA} = \mathrm{Cov(HS)} = 1/4\sigma^2_A \qquad \text{or} \qquad \sigma^2_A = 4\sigma^2_{GCA}$$

Hence,

$$\sigma^2_{AW} = \sigma^2_A - 2\sigma^2_{GCA} = \sigma^2_A - 1/2\sigma^2_A = 1/2\sigma^2_A \qquad \text{or} \qquad \sigma^2_{AW} = 2\sigma^2_{GCA}$$

The individual breeding value defined above (A_{ijklm}) can be predicted using BLUP methodology (written in matrix form):

$$\mathbf{\hat{A} = C'V^{-1}(Y - X\hat{\beta})} \tag{6}$$

where $\mathbf{\hat{A}}$ is the vector of predicted individual breeding values,

C is covariance matrix of **A** and **Y**;

V is the variance matrix of **Y**;

Y and **X** are defined the same as in model (2) and $\mathbf{X\hat{\beta}}$ is BLUE solution of $\mathbf{X\beta}$.

Using (5) we can split **C** into two components:

$$\mathbf{C} = \mathrm{Cov}(\mathbf{A},\mathbf{Y}) = \mathrm{Cov}(\mathbf{G}+\mathbf{A_w},\mathbf{Y}) = \mathrm{Cov}(\mathbf{G},\mathbf{Y}) + \mathrm{Cov}(\mathbf{A_w},\mathbf{Y})$$

Taking a further look at $\mathrm{Cov}(\mathbf{A_w},\mathbf{Y})$, we get

$$\mathrm{Cov}(\mathbf{A_w},\mathbf{Y}) = \mathrm{Cov}(\mathbf{A_w}, \mathbf{G} + \mathbf{A_w} + \mathbf{E}^*)$$

$=\mathrm{Cov}(\mathbf{A_w}, \mathbf{G}) + \mathrm{Cov}(\mathbf{A_w}, \mathbf{A_w}) + \mathrm{Cov}(\mathbf{A_w}, \mathbf{E^*})$

$=\mathbf{0} + \sigma^2_{\mathrm{AW}}\mathbf{I} + \mathbf{0}$

$=2\sigma^2_{\mathrm{GCA}}\mathbf{I}$

Substitute **C** back to (6), we get

$$\begin{aligned}\hat{\mathbf{A}} &= \mathrm{Cov}(\mathbf{G},\mathbf{Y})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}) + \mathrm{Cov}(\mathbf{A_w},\mathbf{Y})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= \hat{\mathbf{G}} + \hat{\mathbf{A}}_{\mathbf{w}}\end{aligned} \qquad (7)$$

where $\hat{\mathbf{G}}$ and $\hat{\mathbf{A}}_{\mathbf{w}}$ are BLUP's of **G** and $\mathbf{A_w}$ respectively.

Since the BLUP solutions of male and female GCA are available from SAS PROC MIXED, we only need to calculate $\hat{\mathbf{A}}_{\mathbf{w}}$ using the following formula:

$$\begin{aligned}\hat{\mathbf{A}}_{\mathbf{w}} &= \mathrm{Cov}(\mathbf{A_w},\mathbf{Y})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= 2\sigma^2_{\mathrm{GCA}}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})\end{aligned} \qquad (8)$$

After obtaining σ^2_{GCA}, $\mathbf{V^{-1}}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$ from SAS PROC MIXED, we can use the above formula to predict the within family breeding value $\mathbf{A_w}$ with SAS PROC IML.

The linear model (Equ. 1) is in fact a general form for full-sib mating designs with similar field designs and the core genetic models (Equ. 3) and (Equ. 4) are true for all full-sib mating designs. So the derived formula (Equ. 8) can be used for other full-sib mating designs than diallel.

Validation of the analytical methods

Validation using simulated data

For generating genetic data, three major population genetic parameters are chosen as the following: narrow sense heritability 0.1, dominance to additive variance ratio 0.3 and type B genetic correlation 0.8. The total phenotypic variance is arbitrarily set to 100. A mating design of a half diallel with 6 parents is used. The field design is a randomized complete block design with 4 tests, 4 replications and 5 trees per family per replication. A 75% imbalance is introduced by applying a survival probability of 0.75 to each individual tree. One thousand diallel data sets were generated for both balanced and unbalanced data and were analyzed using the PROC MIXED method. REML option (restricted maximum likelihood) was used for model fitting in PROC MIXED.

The bias was calculated for both balanced and unbalanced data sets by taking the difference between estimate of parameter and its true value and was expressed as the percentage of true value. Standard deviation was calculated as the square root of sample variance of 1000 variance component estimates. Mean distance was defined as the average difference between estimate and true value, which measured how close the estimates spread around the true value. Either a smaller standard deviation or mean distance indicates a better variance estimator. Both standard deviation and mean distance of the variance estimates were calculated to evaluate the variance component estimates.

Correlation between true genetic value and BLUP served as the criterion of accuracy of genetic gain prediction. Correlation was calculated for 6 parental GCAs, 15 full-sib SCAs and BVs for each simulated data set and then was averaged over 1000 simulations.

Comparison with other methods

Four simulated data sets were used to run diallel analysis with SAS PROC MIXED. The mating design was a 6-parent half-diallel and the field design was a randomized complete block design with 5 test locations, 4 blocks per test and 5 trees per plot. One set is totally balanced and the other three sets have certain imbalance due to missing 5 crosses, 40%

mortality and both missing 5 crosses and 40% mortality. The SAS PROC MIXED was used to run the four data sets, and the estimates were compared with the known parameters used to generate the data.

For comparison, the same data sets were run with the other two computer programs DIALL and GAREML. The statistical validation of these two packages has been previously established (Schaffer and Usanis 1969, Huber 1993). Because DIALL does not allow interactions other than location by genetic effects, the plot effect is incorporated into the error term in the model in order for the results to be comparable among the three programs. The results, including estimates of variance components, BLUP prediction of GCA, SCA from the new method (PROC MIXED), were compared with those from DIALL and GAREML.

Results

Accuracy and bias

Bias, standard deviation and mean distance of variance component estimates

Variance components estimates with SAS PROC MIXED generally agreed with the true values of simulated data. The bias ranged from –4.72% to .13% for all random effects except for SCA by TEST interaction (SCA*TEST, see table 1-1). SCA*TEST has the biggest upward bias (81% for balanced data and 120% for unbalanced data), which may be expected for such a small variance component. The bias increased for all estimates except error variance with imbalance, though this study is not intended to investigate how imbalance would affect accuracy of analysis. For example, bias changed from -1.06% for balanced data sets to -2.84 % at 75% imbalance level for GCA variance, and from -.87% to -1.93% for SCA variance.

The standard deviation was also listed for each random effect in table 1-1. It ranged from .931 to 3.769 and varied with the value of true variance. Except for SCA*TEST, standard deviation is smaller than true variance itself. The degrees of freedom also had an effect on standard deviation in that large degrees of freedom resulted in more accurate variance estimate (e.g. for PLOT and ERROR). Data imbalance also increased standard deviation to certain degree.

Table 1-1. Means, standard deviation and percentage bias of estimated variance components for 1000 simulated data sets

		Balanced data			Unbalanced data (75% survival)		
Effect	True variance	Mean of estimates	S.D.	Bias (%)	Mean of estimates	S.D.	Bias (%)
GCA	5.000	4.947	3.685	-1.06	4.858	3.769	-2.84
SCA	1.500	1.487	1.274	-0.87	1.471	1.347	-1.93
GCA*TEST	1.250	1.219	0.886	-2.48	1.191	0.998	-4.72
SCA*TEST	0.375	0.680	0.931	81.33	0.826	1.153	120.26
PLOT	8.563	8.257	2.311	-3.57	8.169	2.734	-4.60
ERROR	77.062	76.959	3.158	0.13	77.034	3.727	0.04

Table 1-2. Mean distances of variance component estimated from true variance and from sample variance of simulated effect values for 1000 simulated data sets

		Balanced data		Unbalanced data	
Effect	True variance components	M.D. from true variance	M.D. from sample variance	M.D. from true variance	M.D. from sample variance
GCA	5.000	2.968	1.596	2.976	1.695
SCA	1.500	1.015	0.893	1.083	1.000
GCA*TEST	1.250	0.720	0.672	0.808	0.759
SCA*TEST	0.375	0.680	0.682	0.827	0.826
PLOT	8.563	1.858	1.780	2.229	2.137
ERROR	77.062	2.557	1.160	2.974	2.025

Note: M D. =mean distance, i.e. average absolute difference from either true parameter or the sample variance of generated random effects.

Similar to standard deviation, mean distances of variance estimates from either true variance or sample variance of true random effects (true sample variance, as defined and discussed in chapter 2) were listed in table 1-2. The size of mean distance varied with true variance and degrees of freedom (Table 1-2). Overall, its range was from .680 to 2.976. For each random effect, a larger mean distance was observed in unbalanced data than in balanced data. Finally, the mean distance from true sample variance was noticeably smaller than that from true variance for GCA and ERROR.

Correlation between true GCA, SCA and BV and their BLUP predictions

The mean correlation of true genetic value and BLUP prediction was high and very close to .9 for both general combining ability (GCA) and breeding value (GV), and lower (around .6) for specific combining ability (SCA) (see table 1-3).

Standard deviation was small for correlation of both GCA and GV, while a bit higher standard deviation was observed for correlation of SCA. Compared with balanced data, mean correlation for both GCA and GV was lower in unbalanced data. Imbalance also increased standard deviation for both SCA and GV.

Table 1-3. Mean correlation and standard deviation of true and predicted GCA, SCA and GV (BLUP) for 1000 simulated data sets

	Balanced data		Unbalanced data	
Effect	Mean correlation	S.D.	Mean correlation	S.D.
GCA	0.896	0.115	0.896	0.109
SCA	0.606	0.161	0.587	0.167
GV	0.897	0.086	0.885	0.100

Note: BV is obtained by add female and male parental GCA and SCA.

Comparison with GAREML and DIALL

The variance estimates from SAS PROC MIXED, GAREML and DIALL programs were compared (table 1-4). With the balanced data, all three methods produce the exact same estimates of variance components.

Only trivial differences were observed between PROC MIXED and GAREML estimates for three unbalanced data set. Under certain circumstances (e.g. more severe imbalance), slight differences can be observed especially for those estimates close to zero (e.g. SCA, G×E effects), due to the different computing algorithm and iteration criteria of two procedures.

With balanced data, BLUP shrinks all GLS solutions toward zero by a similar degree, hence producing almost the same ranking as GLS (Table 1-5 and 6). Severe imbalance may lead to very different ranking of GCA and SCA effects (Table 1-5 and 1-7). The GCA and SCA predictions were very similar, and the ranks of GCA or SCA were exactly the same between PROC MIXED and GAREML programs in all four data sets (Table 1-5 and 1-6). Theoretically, any differences are due to different variance component estimates, because the same formulas are used for BLUP in both programs. However, with SAS PROC MIXED we can also use other methods to estimate variance components or input predetermined variance component estimates to obtain BLUP solutions.

Table 1-4. Comparison of variance components and individual heritability estimated by SAS PROC MIXED, GAREML and DIALL, with different degree of balance.

	Missing crosses	Survival rate(%)	PROC MIXED	GAREML	DIALL
σ^2_{GCA}	0	100	0.58062	0.58062	0.58062
	0	60	0.50672	0.50667	0.51618
	5	100	0.67712	0.67711	0.73234
	5	60	0.67068	0.67059	0.70840
σ^2_{SCA}	0	100	0.20320	0.20320	0.20320
	0	60	0.14287	0.14312	0.15645
	5	100	0.13655	0.13655	0.13483
	5	60	0.03381	0.03383	0.07237
$\sigma^2_{GCA\times TEST}$	0	100	0.49298	0.49298	0.49298
	0	60	0.68243	0.68243	0.73714
	5	100	0.39800	0.39803	0.40232
	5	60	0.61336	0.61409	0.73606
$\sigma^2_{SCA\times TEST}$	0	100	0.23249	0.23249	0.23249
	0	60	0	0	-0.10363
	5	100	0.07594	0.07594	0.14314
	5	60	0	0	-0.15878
σ^2_{Error}	0	100	9.64388	9.64388	9.64388
	0	60	9.95979	9.95971	10.00917
	5	100	9.78107	9.78106	9.78106
	5	60	10.02713	10.02686	10.12926
h^2	0	100	0.20824	0.20824	0.20824
	0	60	0.17950	0.17948	0.18247
	5	100	0.24470	0.24470	0.26170
	5	60	0.23647	0.23643	0.24667

Table 1-5. Comparison of GCA estimated by SAS PROC MIXED, GAREML and DIALL, with different degree of imbalance.

Parent	PROC MIXED		GAREML		DIALL	
	GCA	Rank	GCA	Rank	GCA	Rank
Balanced						
1	0.64853	1	0.64853	1	0.684	1
2	0.18898	4	0.18898	4	0.199	4
3	0.22858	3	0.22858	3	0.241	3
4	-0.69282	5	-0.69282	5	-0.731	5
5	-0.94262	6	-0.94262	6	-0.995	6
6	0.56935	2	0.56936	2	0.601	2
60% survival						
1	0.54884	2	0.548787	2	0.660	2
2	0.09445	4	0.094454	4	0.018	4
3	0.11460	3	0.114584	3	0.194	3
4	-0.45219	5	-0.452147	5	-0.474	5
5	-0.92773	6	-0.927633	6	-0.972	6
6	0.62203	1	0.621954	1	0.749	1
5 missing crosses						
1	0.80693	1	0.80692	1	1.635	1
2	0.26574	3	0.26574	3	0.113	3
3	-0.11944	4	-0.11944	4	0.021	4
4	-0.64237	5	-0.64237	5	-0.767	5
5	-0.99129	6	-0.99128	6	-1.060	6
6	0.68044	2	0.68044	2	0.615	2
5 missing cross & 60% survival						
1	0.70282	2	0.70267	2	1.919	1
2	0.17964	3	0.17960	3	-0.215	4
3	-0.22298	4	-0.22294	4	0.116	3
4	-0.36576	5	-0.36569	5	-0.516	5
5	-1.07825	6	-1.07803	6	-0.963	6
6	0.78454	1	0.78439	1	0.793	2

Table 1-6. Comparison of SCA estimated by SAS PROC MIXED, GAREML and DIALL for the balanced data set.

	PROC MIXED		GAREML		DIALL	
Cross	SCA	Rank	SCA	Rank	SCA	Rank
1 & 2	0.20831	4	0.20831	4	0.451	4
1 & 3	0.42430	1	0.42430	1	0.823	1
1 & 4	-0.10956	10	-0.10957	10	-0.192	10
1 & 5	-0.49391	15	-0.49391	15	-0.875	15
1 & 6	0.19782	5	0.19783	5	0.477	3
2 & 3	-0.30246	13	-0.30246	13	-0.467	13
2 & 4	0.36020	2	0.36021	2	0.556	2
2 & 5	-0.20323	12	-0.20323	12	-0.433	12
2 & 6	0.00331	8	0.00330	8	0.093	7
3 & 4	-0.47759	14	-0.47759	14	-0.867	14
3 & 5	0.28792	3	0.28792	3	0.409	5
3 & 6	0.14782	6	0.14782	6	0.343	6
4 & 5	0.10676	7	0.10676	7	-0.006	8
4 & 6	-0.12227	11	-0.12227	11	-0.222	11
5 & 6	-0.02742	9	-0.02742	9	-0.090	9

Table 1-7. Comparison of SCA estimated by SAS PROC MIXED, GAREML and DIALL, for the unbalanced data set with 5 missing crosses and 60% survival

	PROC MIXED		GAREML		DIALL	
Cross	GCA	Rank	GCA	Rank	GCA	Rank
1 & 2	0.02506	3	0.02508	3	-0.091	7
1 & 3	-		-		-	
1 & 4	-		-		-	
1 & 5	-		-		-	
1 & 6	0.01037	5	0.01038	5	-0.726	10
2 & 3	-0.03848	8	-0.03850	8	-0.047	6
2 & 4	0.06740	2	0.06744	2	1.026	1
2 & 5	-0.04492	9	-0.04495	9	0.051	4
2 & 6	-		-		-	
3 & 4	-0.06740	10	-0.06623	10	-0.525	9
3 & 5	-		-		-	
3 & 6	0.09344	1	0.09349	1	0.493	2
4 & 5	0.01759	4	0.01760	4	0.036	5
4 & 6	-0.03723	7	-0.03725	7	0.140	3
5 & 6	-0.02702	6	-0.02703	6	-0.210	8

Discussion

The SAS PROC MIXED procedure can be modified to analyze genetic data from a half-diallel mating design. A dummy variable approach was used to overcome the technical problem presented by the unique feature of diallel designs. The mixed model framework produces all needed genetic parameter estimates and predictions for GCA and SCA.

Through simulation, we have shown that this new procedure can provide accurate variance component estimates for all random variables as compared with the true vales. Since REML does not allow negative estimates of variance components and sets negative estimates to zero, for a small true variance, upward bias is observed and expected. In this simulation study, SCA*TEST only accounts for 0.375% of the total variance and is seriously biased (Table 1-1). Fortunately, for most breeding programs, this effect is only of trivial interest and even eliminated from the model to reduce computational complexity (Lu et al, 1999). For other variance components, REML estimates tend to under estimate the true value due to the upper bias for SCA*TEST, but the biases are small (less than 5% for all). The bias increases with unbalanced data sets. Overall, these results are typical for REML estimates.

Correlations between true genetic values and BLUP estimates are high. Considering the small number of parents of 6, correlation for GCA and BV is high for both balanced and unbalanced data. Small standard deviation for correlation of both GCA and GV indicated that consistently high correlations were achieved in majority of simulated data sets. Lower correlation with larger standard deviation for SCA is expected because of low dominance variance. Although the mating design and field design may not be the best for variance component estimates, the relatively accurate estimates and high correlations with true value for GCA, SCA and GV predictions should be reliable for ranking family and selection.

Under a balanced situation, the SAS MIXED procedure produces similar variance estimates to GAREML and DIALL. With the increase of imbalance, the difference between REML estimates (both PROC MIXED and GAREML) and ANOVA based estimates (DIALL) becomes distinct. The REML estimator was shown to be superior to the ANOVA based estimator in simulation studies (Swallow and Monahan 1984; Searle et al. 1992; Huber, 1993).

Compared with GAREML and DIALL, one of the advantages offered by the SAS MIXED procedure is that it can be used with model fitting technologies (ML and MIVQUE0) and modified by changing the starting point of iteration, maximum number of iterations and convergence criteria etc. With any degree of imbalance in the data, the SAS PROC MIXED should have better estimates because it uses MIVQUE0 estimators as starting point of iteration and has either stricter default convergence criteria. In fact, when differences do exist, –2 REML log likelihood statistics for GAREML estimates, which can be calculated in PROC MIXED by inputting variance components from GAREML estimates, are slightly larger than those from SAS PROC MIXED. The detailed information about SAS PROC MIXED is described in SAS/STAT system manuals and related publications (SAS Inst. Inc., 1996; Littell et al, 1996).

DIALL calculates general least square solutions (GLS) of genetic effects by treating them as fixed effects and adding sum-to-zero restriction, while PROC MIXED and GAREML provide a consistent treatment of GCA's and SCA's as random effects and produce their best linear unbiased predictions (BLUP) (Huber, 1992). The desirable properties of BLUP include maximizing the probability of obtaining correct parental rankings from the data and minimizing the error associated with using the parental values obtained in future applications (Huber, 1993; White and Hodge, 1988). BLUP methodology is particularly more desirable when data are unbalanced and absolute gain assessment is needed.

Under unbalanced data, the proposed procedure provides superior estimates of variance components and more desirable predictions (BLUP) of genetic effects compared with DIALL. The difference from GAREML is small. An additional advantage of using SAS is that PROC MIXED can output any information into data sets for further manipulation and calculation. For example, the fixed effects can be added back to GCA and SCA estimates to calculate genetic gains or for other purposes. For disconnected diallel analysis, the diallel effect can be estimated and used to adjust for GCA and SCA estimates. PROC MIXED will also provide other useful information in its outputs. In addition, this practical procedure can be modified to estimate ordinary least square (OLS) or generalized least square (GLS) solutions for GCA and SCA to save computing time. It also provides a friendly interface and many other flexible options compared with other specific packages.

Another useful application is to predict individual tree breeding values using the formula derived in this paper. Usually, within family heritability and within family deviation (i.e. $\mathrm{Y}_{\mathrm{ijklm}} - \overline{\mathrm{Y}}_{..\mathrm{kl.}}$) are used to calculate within family breeding values. The advantage of formula (8) is that $\hat{\mathbf{A}}_{\mathbf{w}}$ has desirable properties of BLUP and is particularly more appropriate with unbalanced data. All the information that is needed for calculation can be obtained from PROC MIXED. By first using PROC MIXED to output variance component estimates, inverse of variance matrix and BLUE of fixed effects, we can easily predict individual tree breeding values using IML once again.

This procedure can further be developed to deal with more complex situations, such as disconnected half-diallel sets and multiple test series. Such complications can be accommodated in the mixed model, so only some modifications are needed in constructing dummy variables.

Reference

[1] Borralho, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. In: Eucalypt plantations: improving fibre yield and quality. Proceedings of CRCTHF-IUFRO Conference. Hobart, Australia. p141-145.

[2] Dean, C.A. and Correll, R.L. Analysis of diallel matings with missing values. *Silvae Genetica* 37(5-6):187-197. 1988.

[3] Huber, D.A. et al, Ordinary least squares estimation of general and specific combining abilities from half-diallel mating designs. *Silvae Genetica* 41(4-5): 263-273 1992.

[4] Huber, D.A. Optimal mating designs and optimal tehcniques for analysis of quantitative traits in forest genetics. Ph. D. dissertation. University of Florida. 1993

[5] Li, B., McKeand, S.E., Hatcher, A.V., and Weir, R.J. Genetic gains of second generation selections from the NCSU-Industry Cooperative Tree Improvement Program. Pro. 24th South. For. Tree Impr. Conf., p. 234-238. Orlando, Florida, June 9-12, 1997.

[1] Li, B., McKeand, S.E., and Weir, R.J. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the south-eastern US. p. 164-68. In: Tree improvement for sustainable tropical forestry. Proceedings of QFRI-IUFRO Conference. Caloundra, Queensland, Australia. 1996.

[6] Littell, R.C. et al. SAS system for mixed models. SAS Institute Inc. Cary, NC. 1996

[7] Lu, P.X., Huber, D.A., and White, T.L. Potential biases of incomplete linear models in heritability estimation and breeding value prediction. Can. J. For. Res. 29: 724-736. 1999.

[8] SAS Institute Inc. SAS/STAT Software: Changes and enhancements (through release 6.11). Cary, NC. 1996.

[9] Schaffer, H.G. & Usanis, R.A. General least square analysis of diallel experiments. A computer program DIALL. N.C. State Univ., Genet. Dept., Res., Rep. 1. 1969.

[10]Searle, S.R., Casella, G., and McCulloch, C.E. Variance components. John Wiley & Sons Inc., New York, 501p. 1992.

[11]Swallow, W.H. and Monahan, J.F. Monte Carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components. Technometrics 26(1): 47-57. 1984.

[12]White T. L. and Hodge, G. R. Best linear prediction of breeding values in a forest tree improvement program. Theor. Appl. Genet. 76:719-727. 1988.

[13]Yanchuk, A. D. General and specific combining ability from disconnected partial diallels of coastal douglas-fir. *Silvae Genetica* 45(1):37-45 1996.

[14]Zobel, B.J. & Talbert, J. Applied forest tree improvement. John Wiley and Sons. New York. NY. 1984.

Chapter 2 Optimal analytical methodology for disconnected diallel test

Introduction

In tree breeding disconnected half-diallel mating design has been widely used for progeny testing and regenerating a breeding population (Yanchuk, 1996; Huber, 1994). While disconnected diallel design has the practical advantage over one large diallel in that far fewer crosses per parent are needed and mating can be completed within a relatively short time, the analysis of disconnected design is more complicated and controversial. A common question concerns the choice of analytical linear models, whether and how we should include a diallel effect in the analysis. When a diallel effect is included, should it be random or fixed? Another important issue is how we compare or rank parents and full-sib families from disconnected diallels in different test series. This study is to examine analytical models for analyzing disconnected diallel design. The results of this study can also be generalized to any disconnected mating designs. The theoretical basis of diallel effect will be evaluated and then computer simulated data will be used to evaluate three different analytical models. In the second part of this chapter, we will compare different analytical methods for dealing with disconnected diallels from multiple test series and try to identify the best one.

I. Comparison of alternative models in analysis of disconnected half diallel mating design

Abstract

Simulated data generated with known parameters were analyzed using Best Linear Unbiased Prediction (BLUP) methodology to compare three alternative models, which include diallel as fixed (Model 1), random effect (Model 2) or no diallel effect (Model 3). Both Model 1 and Model 3 produced unbiased GCA variance estimate. Model 2 resulted in downward biased estimate of GCA variance component and occasionally produced unrealistically large diallel variance estimate. Model 3 was slightly better than the other two models in accuracy of GCA variance estimates, while the difference in accuracy between Model 1 and Model 2 was rather trivial. The accuracy of BLUP prediction for three models, measured as the correlation between true genetic value and prediction, was very close, with Model 3 slightly better than the other two. Model 3 is preferred under Randomized Complete Block Design (RCBD) and random selection of parents, while the Model 1 performed well overall in both variance component estimation and BLUP. Model 2 is not recommended in most situations because of its undesirable GCA variance estimate.

Key words: Diallel mating design, General combining ability (GCA), Breeding value (BV), Best linear unbiased prediction (BLUP), Linear Mixed model

Intoduction

Disconnected diallel designs have been widely used in crop and tree breeding programs (Yanchuk, 1996; Huber, 1994). A set of partial diallels could be tested under same environmental condition, or they could be planted at different sites. Variance component estimation is often carried out on each separate diallel set and then averaged over the whole test series (Li et al, 1996). With the advance in computing speed and model fitting technologies, combined analysis for entire test series has become feasible. It would eliminate any problems due to averaging variance components across disconnected diallel sets. In case of disconnected diallels tested at the same site where heterogeneity does not present a problem, it is more desirable to use combined analysis. However, for the combined analysis a critical question is how to deal with the diallel effect in the linear model. It could be considered as sampling error and treated as random effect. It could also be conveniently treated as fixed effect to assess the differences among diallels and then be used to adjust for gain prediction later. It could also be totally ignored in the model. There has been no information in the literature to address these questions: How do these different treatments of diallel effect affect the analysis of genetic variance components and breeding value prediction? What is the best treatment of diallel effect in the analytical model? Which model is the best for variance components estimation or genetic effect prediction? This study is designed to address these questions with simulated data. Three linear models with diallel treated as random, fixed and no diallel effect, will be evaluated for their merits for estimating variance components and predicting family breeding value.

Theoretical background

Let 12 parents represent a random sample from a base population, assuming general combining ability (GCA) effect of any parent distributed as N(0, σ^2_g) and specific combining ability (SCA) of any mating between two parents distributed as N(0, σ^2_s). They are randomly assigned into a two 6-parent diallels mating design and are tested under the same environmental condition. Let GCA and SCA be $G_{i(o)}$ and $S_{ij(o)}$, *i*, *j*=1 to 6, *o*=1,2. From theoretical point of view, there should be no genetic difference other than random sampling error between the two diallel sets.

If diallel effect D_o (*o*=1,2) is included in the model to assess the random sampling drift of genetic variance, the assumption of independence of diallel effect and GCA effects $G_{i(o)}$ will be violated. In other words, there is no sampling process that leads to a model, which includes an independently identically distributed (iid) nonzero diallel effect and an iid GCA effect. If the genetic diallel effect D_o (*o*=1,2) does exist as NID(0, σ^2_d), crosses of the same diallel would have a common genetic effect D_o. Then the total genetic value $D_o+G_{i(o)}+G_{j(o)}+S_{ij(o)}$ would not be independent within a particular diallel. The covariance between genetic values of two crosses within the same diallel is σ^2_d. On the other hand, if parents are randomly assigned into 2 diallels, a particular cross will have the same probability of appearing in both diallels. On the population level, the true genetic value of a cross should not change whether it is in diallel 1 or diallel 2. In other words, the diallel or the mating of a set of parents is the result of random sampling, which would not produce any genetic effect.

In addition, as in the case of diallel tests from the second generation NCSU loblolly pine tree improvement program, no environmental difference in randomized complete block design (RCBD) is expected between 2 diallels, since trees of 2 diallels are randomly assigned within each block in each test. In this case, no specific causal factor can be attributed to diallel effect. Under this scenario, we may simply set variance of diallel (σ_d^2) equal to zero or eliminate this effect from the model.

On the other hand, if either random selection of parents is violated or some unknown environmental factors are confounded with diallel effect due to flaws in the field experimental design, then elimination of diallel effect will result in an inflated estimate of GCA variance. A complete model with diallel effect may be needed to deal with both situations. However, estimation of an additional parameter (σ_d^2 if treated as random or d if treated as fixed) could potentially affect estimation of other parameters. If diallel is treated as a random effect, it could cause bias in variance estimation using the variance estimation technology REML as in SAS MIXED procedure or GAREML. Since negative variance estimates are not allowed in REML, even if σ_d^2=0 the expected value of REML estimate of σ_d^2 is greater than zero. To worsen the situation, two levels of diallel effect will give an inaccurate estimate of σ_d^2. The variation of this estimate could be fairly large.

In this part of the study, we will evaluate the following 3 models using simulation:

1. Complete model with diallel as fixed effect (fixed model or Model 1);
2. Complete model with diallel as random effect (random model or Model 2);
3. Incomplete model with no diallel effect (incomplete model or Model 3).

Simulated data with two and four disconnected diallels will be examimed for all three models.

Method

Linear model

The linear model used in simulation follows a common scalar linear model for a disconnected diallel mating design with a randomized complete block design (RCBD) at multiple test locations. The mating design consists of 2 or 4 disconnected 6-parent half diallels. Field design is RCBD with 4 tests, 6 blocks and 6 trees per cross per block in each test. The linear model is the following.

$$Y_{ijoklm} = \mu + T_i + B_{j(i)} + D_o + G_{k(o)} + G_{l(o)} + S_{kl(o)} + TG_{ik(o)} + TG_{il(o)} + TS_{ikl(o)} + P_{ijokl} + E_{ijoklm} \qquad (1)$$

where,

Y_{ijoklm} is the *m*th observation of the *j*th block within *i*th test for the *kl*th cross of the *o*th diallel;

μ is the overall mean;

T_i is the fixed effect of *i*th test, i=1 to 4;

$B_{j(i)}$ is the fixed effect of *j*th block within *i*th test, j=1 to 6;

D_o is the *o*th diallel effect, o=1 to d, random in Model 1, fixed in Model 2, or none in Model 3;

$G_{k(o)}$ or $G_{l(o)}$ is the GCA effect of *k*th female or *l*th male of *o*th diallel (k,l=1, … , 6; k<l) ~ NID(0, σ^2_{GCA});

$S_{kl(o)}$ is the SCA effect of *k*th and *l*th parents of *o*th diallel ~ NID(0, σ^2_{SCA});

$TG_{ik(o)}$ or $TG_{il(o)}$ is the *i*th test by *k*th female or *l*th male GCA of *o*th diallel interaction ~ NID(0, $\sigma^2_{TEST*GCA}$);

$TS_{ikl(o)}$ is the *i*th test by of *k*th and *l*th parents SCA of *o*th diallel interaction ~ NID(0, $\sigma^2_{TEST*SCA}$);

P_{ijokl} is the random effect of plot for the *kl*th cross of *o*th diallel in the *j*th block within *i*th test ~ NID(0, σ^2_{PLOT});

E_{ijoklm} is the within plot error term ~ NID(0, σ^2_E).

Plot means were used in simulation to reduce computational time without affecting results as long as the number of trees per plot are uniform. Plot mean analysis also can deal with imbalance due to missing plots and crosses. Hence the model reduces to the following:

$$\overline{Y}_{ijokl.} = \mu + T_i + B_{j(i)} + D_o + G_{k(o)} + G_{l(o)} + S_{kl(o)} + TG_{ik(o)} + TG_{il(o)} + TS_{ikl(o)} + PT_{ijokl} \qquad (2)$$

where, $\overline{Y}_{ijokl.}$ is the plot mean observation of the *j*th block within *i*th test for the *kl*th cross of *o*th diallel;

PT_{ijokl} is the random plot mean error for the *kl*th cross of *o*th diallel in the *j*th block within ith test ~ NID(0, σ^2_{PT}), equivalent to $P_{ijokl}+\overline{E}_{ijokl.}$ in equation (1) and hence $\sigma^2_{PT}=\sigma^2_{PLOT}+1/6\sigma^2_E$;

All other terms are defined the same as in equation (1).

Generation of simulated diallel data sets

Three important parameters are considered in generating diallel data sets: narrow sense heritability, h^2; dominance to additive genetic variance ratio, γ; type B genetic correlation, r_B. For each parameter, two levels are chosen to represent a higher and lower value within the normal range of that parameter, resulting in a total of 8 levels of genetic determinations as listed in table 2.1.1.

Without losing generality, the total phenotypic variation is set to be 100. With different parameter combinations, values of σ^2_g, σ^2_s, σ^2_{gt}, and σ^2_{st} were calculated. σ^2_p is assumed to account for 10% of the rest of the variation, thus gives the solutions for σ^2_p, σ^2_e. The plot mean variance σ^2_{pt} is $\sigma^2_p+(1/6)\sigma^2_e$. σ^2_d is set to zero, as the assumption of random selection of parents and RCBD field design hold.

Table 2.1.1. Parameter structure for generating simulated data sets

Label	h^2	γ	r_B	σ^2_g	σ^2_s	σ^2_{gt}	σ^2_{st}	σ^2_p	σ^2_e	σ^2_{pt}
1	.1	.3	.6	2.500	.750	1.667	.500	9.042	81.374	22.604
2	.1	.3	.8	2.500	.750	.625	.188	9.281	83.531	23.203
3	.1	.7	.6	2.500	1.750	1.667	1.167	8.875	79.874	22.187
4	.1	.7	.8	2.500	1.750	.625	.438	9.156	82.406	22.890
5	.2	.3	.6	5.000	1.500	3.333	1.00	8.083	72.751	20.208
6	.2	.3	.8	5.000	1.500	1.250	.375	8.563	77.062	21.407
7	.2	.7	.6	5.000	3.500	3.333	2.333	7.750	69.751	19.375
8	.2	.7	.8	5.000	3.500	1.250	.875	8.313	74.812	20.782

In simulation, test and block effects were generated as random effects, so that each simulated test series has a different set of fixed effects. In the real data, the average test-to -test variation is about 80% of phenotypic variance and block within test variation is 10%. Hence, the variance of test and block are set to be 80 and 10, respectively.

SAS IML procedure was used to produce diallel data sets with the above known parameters. More specifically, by using random number generating function RANOR and multiplying by the square root of variance component we obtained a random vector for each random effect. Random deviations from each vector are then summed up according to the linear model to produce the plot mean vector. The random vector for GCA and SCA effects were also outputted from IML so that the exact genetic values of each parent and cross were known.

For parameter combination 2 and 6, which are the most common genetic structures in practical tree breeding, two types of imbalance at plot level were considered: missing crosses and missing plots. Non-uniform plot survival of different tests and crosses were also taken into account. Altogether four levels of imbalance were generated for these two parameter structures (table 2.1.2).

Table 2.1.2. Imbalance levels considered in simulation study

Label	Imbalance description
a	20% missing plot
b	5 missing crosses per diallel
c	20% missing plot plus 5 missing crosses per diallel
d	15% missing plot with 1 low survival test and 5 low survival crosses

PROC MIXED analysis

For each test series, 1000 simulated data sets were generated and analyzed using the SAS MIXED procedure for each model (as described in chapter 1). The default model fitting method REML was used since it has desirable properties in many simulation studies and is widely used in practice. For Model 1, a fixed diallel effect was included in the

MODEL statement. For Model 2, a random diallel effect was included in RANDOM statement. No diallel effect was included in either the MODEL or the RANDOM statement for Model 3.

For each model, best linear unbiased prediction of GCA and SCA effect was obtained for each parent and cross by outputting solution of random effects from PROC MIXED. In the complete model (Model 1 and 2), adjusted GCA was also calculated by adding half of diallel effect (random or fixed) to the parental GCA, assuming both parents of each cross contributed equally to the diallel effect. This adjustment will make parents in two diallels more comparable (Huber et al, 1992).

$$GCA_{adj}=GCA+0.5*d$$

Genetic value of each full-sib family was then calculated by adding two parental GCA's and SCA of the cross.

$$GV=GCA_i+GCA_j+SCA_{ij}$$

Thus, three sets of variance component estimates and BLUP of GCA and SCA were obtained for each test series.

Criteria of comparison

Several criteria were used to compare these three linear models.

1. Bias of variance component estimate from true population variance and true sample variance;
2. Mean distance of variance component estimates from true population variance and sample variance;
3. Mean square error of variance component estimates from true parameter and sample variance;
4. Correlation between variance component estimates and true sample variances;
5. Mean distance of GCA and breeding value prediction from true genetic values;
6. Correlation between BLUP's (of GCA or GV) and true genetic values.

"True sample variance" of a random effect is defined as the variance of a given random sample of that random effect. Its expectation is just the population variance, which is unknown in practical analysis. But it can be calculated in this simulated study since the true values of random effects are known. For example, 12 true GCA values $G_{1,\ldots,}G_{12}$ are generated in each run of simulation, the true sample variance S_g^2 is just the sample variance of these 12 values.

$$S_g^2 = \frac{\sum_{i=1}^{12}(g_i - \overline{g}_.)^2}{11}$$

For a given random sample of GCA effect, the goal of any variance component estimation method is to estimate this sample variance S_g^2. Hence given S_g^2, an unbiased estimator should have the conditional expectation S_g^2, which justified its use as the unbiasness criteria. Moreover, any estimator cannot be better than S_g^2 itself (if S_g^2 is known), because errors always exist in any practical analysis. The advantages of using true sample variance are: 1) In a small simulation study, mean of S_g^2 may not converge closely enough to the true population variance. As a result, sampling error rather than analytical method causes the bias for the most part. The use of S_g^2 eliminates the bias due to sampling error; 2) Correlation of S_g^2 and variance estimates can be further calculated to evaluate the different analytical methods. The higher is the correlation, the better the method; 3) The variance of variance estimates can be decomposed into the part due to sampling drift i.e. the change of S_g^2 and the part due to estimation error.

Results and discussion

Variance components estimation

The variance of diallel effect in the random model

Although there were no true diallel effects in simulated data sets, its estimated variance from the RANDOM model (Model 2) was not negligible in all parameter combinations (Table 2.1.3). The distribution of diallel variance estimate was extremely skewed toward the left. As a result, more than half of the estimates fall to zero. But nonzero estimates frequently occurred regardless of the parameter structure.

Table 2.1.3. Summary statistics of σ_d^2 estimate in Model 2 from analysis of simulated data sets (1000 for each parameter structure and imbalance level), including mean, standard deviation (S.D.), Coefficient of Variation (CV), minimum (MIN), maximum (MAX) and 75 percentile

Parameter Structure and imbalance level	Mean	S. D.	CV(%)	MIN	MAX	75 percentile
1	1.0534	2.3716	225.1	0	17.122	0.8216
2	1.0855	2.4605	226.7	0	23.383	0.9786
3	1.1693	2.6129	223.5	0	20.584	1.0693
4	0.9688	2.3042	237.8	0	21.796	0.7248
5	1.9212	4.4402	231.1	0	37.816	1.3080
6	1.7921	3.9988	223.1	0	30.479	1.4802
7	1.9459	4.4657	229.5	0	31.761	1.6591
8	1.6936	3.8443	227.0	0	32.535	1.1849
2a	0.9172	2.2891	249.6	0	25.1206	0.6007
2b	0.8902	2.1068	236.7	0	20.477	0.7264
2c	1.0435	2.3364	223.9	0	20.21	0.9468
2d	1.0265	2.371	231.0	0	18.346	0.8591
6a	2.0366	4.807	236.0	0	41.564	1.5866
6b	1.8549	4.7692	257.1	0	54.587	1.2971
6c	2.0549	4.5364	220.8	0	38.091	1.7587
6d	1.7835	4.348	243.8	0	35.639	1.2121

Note: * — parameter structure and imbalance label, see table 2.1.1, 2

The mean variance was 1.02, accounting for 41% of σ_g^2 for h^2=0.1, and 1.89 or 38% of σ_g^2 for h^2=0.2. The variation of variance estimates was huge with coefficient of variance (C.V.) exceeding 220% in all cases. As also listed in Table 2.1.3, 75 percentiles indicate that 25% of estimates were fairly large, which, on average, was at least 34% of σ_g^2 when h^2=0.1 and 29% of σ_g^2 when h^2=0.2. Moreover, in all parameter structures simulation occasionally produced extreme estimates, which were 10 times the magnitude of σ_g^2.

GCA Variance estimates

Bias

Two measures of bias of GCA variance estimates are listed in Table 2.1.4. All values are expressed as the percentage of true values. As we can see, both Model 1 and 3 provided unbiased estimate of σ_g^2. Averaged across 16 simulation runs for data sets, the bias was less than 0.5% for both models. But the GCA variance estimate from RANDOM model (Model 2) was biased downward consistently over different simulations. The average bias was –6.03% from population variance and –6.37% from true sample variance.

No further bias seems to be introduced due to imbalance. We also notice that true sample variance (the 3 columns on the right, Table 2.1.4) criteria eliminates the random drift effect and hence gives a clearer indication of both the trend and the degree of bias.

Table 2.1.4. Percentage bias of σ_g^2 estimate in 3 models from analysis of simulated data sets (1000 for each parameter structure and imbalance level)

	Percentage Bias					
PN*	From true population variance			From true sample variance		
	MODEL1	MODEL2	MODEL3	MODEL1	MODEL2	MODEL3
1	2.38	-4.53	2.20	2.76	-4.18	2.57
2	-0.12	-6.50	0.48	-0.55	-6.90	0.05
3	0.96	-7.00	1.09	-1.10	-8.87	-0.98
4	-3.95	-10.08	-3.98	-2.58	-9.52	-2.61
5	3.85	-2.75	3.05	2.14	-4.35	1.35
6	1.63	-4.19	1.32	0.60	-5.16	0.29
7	0.35	-7.07	-0.60	0.71	-6.74	-0.25
8	1.91	-4.95	0.75	1.48	-5.40	0.33
2a	-2.10	-7.88	1.58	0.48	-6.65	-0.27
2b	2.38	-4.57	1.44	1.54	-5.36	0.60
2c	-0.62	-7.57	0.36	0.71	-6.33	0.97
2d	3.37	-3.48	3.35	0.53	-6.13	0.51
6a	0.13	-6.13	0.33	-0.31	-6.54	-0.11
6b	-0.98	-7.01	-1.21	-0.28	-6.35	-0.51
6c	-0.14	-6.28	0.37	-1.80	-7.85	-1.30
6d	-0.39	-6.55	-0.94	0.58	-5.65	0.02
Mean	0.54	-6.03	0.60	0.31	-6.37	0.04

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

Variance of $\hat{\sigma}_g^2$ and correlation between S_g^2 and $\hat{\sigma}_g^2$

In order for variance of estimates to be comparable for different population variances, C.V. (coefficient of variation) instead of variance was used (Table 2.1.5). The correlation between true sample variance S_g^2 and variance estimate $\hat{\sigma}_g^2$ is also listed in the Table 2.1.5.

No significant difference in C.V. was observed among the three models. However, Model 3 consistently had less variation than other two models. Although in most cases Model 1 performed a bit better than Model 2, but the difference was rather trivial.

In terms of correlation between S_g^2 and $\hat{\sigma}_g^2$, the correlation in model 3 was the highest i.e. .02~.05 higher than that of Model 2, which in turn was about .01~.02 higher than that of Model 1.

The effect of parameter structure on variance estimate can be seen here. High heritability increased correlation between S_g^2 and $\hat{\sigma}_g^2$ from .7064 for h^2=.1 to .7470 for h^2=.2 when γ=.3 and r_B=.6. Other changes in parameter structure such as low dominance genetic control and high type B genetic correlation can also reduce variation of variance estimate and increase correlation b between S^2_g and $\hat{\sigma}_g^2$.

Table 2.1.5. Coefficient of variation (C.V.) of σ^2_g estimate and correlation between S^2_g and σ^2_g estimate in 3 models from analysis of simulated data sets (1000 for each parameter structure and imbalance level)

PN*	C. V.			Corr(S_g^2, $\hat{\sigma}_g^2$)		
	Model1	Model2	Model3	Model1	Model2	Model3
1	61.42	61.90	57.90	0.7064	0.7160	0.7484
2	56.76	56.70	52.93	0.7553	0.7700	0.8146
3	67.30	68.67	63.80	0.6718	0.6819	0.7108
4	61.71	62.65	58.79	0.7007	0.7164	0.7610
5	57.78	57.95	55.34	0.7470	0.7591	0.7824
6	52.76	52.95	50.55	0.7966	0.8107	0.8412
7	63.84	64.96	60.77	0.6788	0.6897	0.7125
8	59.16	59.42	55.58	0.7265	0.7381	0.7594
2a	59.78	60.69	56.97	0.7456	0.7572	0.7913
2b	55.94	52.23	52.22	0.7461	0.7584	0.7890
2c	59.30	59.64	55.65	0.6970	0.7070	0.7482
2d	55.21	55.04	51.49	0.7235	0.7387	0.7770
6a	55.95	56.07	53.12	0.7738	0.7884	0.8181
6b	55.41	55.33	52.97	0.8006	0.8119	0.8506
6c	56.61	56.74	53.57	0.7475	0.7624	0.8014
6d	54.40	54.02	51.10	0.7650	0.7833	0.8248
4D6	37.01	35.93	33.99	0.7692	0.7986	0.8493

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

Mean distance and mean square error

The mean distance (MD) and mean square error (MSE) provide additional measurements of how estimates are scattered around the target value (as shown in table 2.1.6 and 2.1.7). The less MD or MSE is, the better the estimate is.

Table 2.1.6. Mean distance of $\sigma^2{}_g$ estimate in 3 models from analysis of simulated data sets (1000 for each parameter structure and imbalance level)

		From σ_g^2			From S_g^2	
PN*	MODEL1	MODEL2	MODEL3	MODEL1	MODEL2	MODEL3
1	1.2236	1.1794	1.1514	0.8709	0.8232	0.7781
2	1.1217	1.0763	1.0518	0.7284	0.6808	0.6060
3	1.3354	1.2912	1.2637	0.9904	0.9446	0.8901
4	1.1859	1.1516	1.1290	0.8211	0.7835	0.7222
2a	1.1750	1.1382	1.1124	0.7681	0.7289	0.6783
2b	1.1174	1.0637	1.0289	0.7571	0.7046	0.6430
2c	1.1583	1.1179	1.1064	0.8235	0.7743	0.7187
2d	1.1167	1.0572	1.0488	0.7835	0.7312	0.6658
5	2.3356	2.3393	2.2295	1.5550	1.4579	1.3954
6	2.0845	2.0222	2.0138	1.2347	1.1613	1.0686
7	2.5424	2.4508	2.4050	1.8614	1.7623	1.6702
8	2.3828	2.2735	2.2129	1.6383	1.5392	1.4448
6a	2.2139	2.1393	2.1275	1.3838	1.2944	1.1899
6b	2.1908	2.1173	2.0926	1.2642	1.1882	1.0834
6c	2.2112	2.1412	2.1197	1.4512	1.3867	1.2627
6d	2.1315	2.0473	2.0184	1.3267	1.2255	1.0811
4D6	1.4616	1.3971	1.3654	0.9126	0.8125	0.7094

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

Both true population variance σ_g^2 and true sample variance S_g^2 were used as target values. MD and MSE calculated from S_g^2 are significantly less than those from σ_g^2, consistently over all parameter structure and imbalance level. The difference in MD calculated from two criteria (S_g^2 and σ_g^2) was about .4 for h^2=.1 and .8 for h^2=.2, which roughly accounted for one third reduction in MD from σ_g^2. Unlike MD, total MSE can be decomposed into the part due to sampling drift and the part due to estimation error. The latter can be measured by MSE from true sample variance. The difference in MSE calculated from two target values (S_g^2 and σ_g^2) is significant and accounted for about a half of total MSE from true population variance σ_g^2. Hence about half of total MSE was caused by sampling drift and was removed from comparing criterion. Because of this reduction in MD or MSE, the differences among three models were more distinguishable for criteria based on true sample variance S_g^2.

Model 3 had the smallest MD and MSE from S_g^2 and σ_g^2. MD and MSE were larger for both Model 2 and Model 1. Moreover Model 1 had slightly larger MD and MSE than Model 2. These conclusions hold regardless of parameter structures and data imbalance levels. Therefore, judging only by MD and MSE, Model 3 is the best choice, followed by model 2 and model 1.

Similar effects of parameter structure were shown on estimate. Either low dominance variance or low genetic by environmental effect can improve the accuracy of GCA variance estimation. The effect of heritability or σ_g^2 is not straightforward here because both MD and MSE will change when the magnitude of the true parameter σ_g^2 or sample variance S_g^2 changes, as shown in the table. The differences in both MD and MSE were not distinguishable among different imbalance levels in this study. However, the increase in MSE was evident in unbalanced data set, compared with results from balanced data sets (Table 2.1.7).

Table 2.1.7. Mean square error of σ^2_g estimate in 3 models from analysis of simulated data sets (1000 for each parameter structure and imbalance level)

	From σ_g^2			From S_g^2		
PN*	MODEL1	MODEL2	MODEL3	MODEL1	MODEL2	MODEL3
1	2.4726	2.1932	2.1893	1.2428	1.0738	0.9668
2	2.0064	1.814	1.766	0.8622	0.7478	0.5941
3	2.8809	2.5772	2.5957	1.5853	1.4138	1.2892
4	2.2035	2.0227	1.9993	1.122	1.0095	0.8415
2a	2.1945	1.9911	1.965	0.9761	0.8599	0.736
2b	2.0450	1.8139	1.7511	0.9089	0.7846	0.6610
2c	2.1668	1.9335	1.9199	1.1145	0.9767	0.8457
2d	2.0367	1.7675	1.7732	0.9687	0.8318	0.7011
5	9.0209	7.9529	8.1363	3.9871	3.4122	3.1544
6	7.1879	6.4704	6.5558	2.6254	2.2816	1.916
7	10.2509	9.227	9.1131	5.5319	4.8872	4.4873
8	9.0767	8.0291	7.8253	4.3014	3.701	3.315
6a	7.8376	7.0140	7.0924	3.1451	2.7382	2.3459
6b	7.5134	6.7357	6.8368	2.6969	2.3632	1.8907
6c	7.9817	7.1604	7.2199	3.5312	3.1185	2.5928
6d	7.3252	6.4707	6.3942	3.0398	2.5528	2.0433
4D6	3.3446	2.9287	2.8458	1.3658	1.0918	0.7947

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

GCA and GV prediction

Mean distance and mean square error

The mean distance and mean square error of parental GCA for all simulations are listed in table 2.1.8. Using the diallel effect adjustment, we see significant improvement in GCA prediction in Model 1, while less improvement was observed in the random model (Model 2). For all parameter structures and imbalance levels, the difference of mean distance between two adjusted GCAs was very small, with both slightly larger than GCA prediction in Model 3.

While the mean distance and mean square error of full-sib family genetic value (GV) were larger than those of GCA (as shown in table 2.1.9), similar results were found when different models were compared. Diallel effect adjustment resulted in significant improvement in GCA prediction in both Model 1 and Model 2, though less improvement was observed in Model 2. Regardless of parameter structures and imbalance levels, only trivial difference was found in mean distance and mean square error between two adjusted GVs (in Model 1 and Model 2). Both Model 1 and Model 2 had slightly larger mean distance and mean square error in GV prediction than those in Model 3.

The parameter structures used in this study affected GCA or GV predictions in a similar way to GCA variance estimates. Either low dominance genetic control or high type B genetic correlation could improve the accuracy of prediction. Both MD and MSE increased as the true parameter σ_g^2 or heritability increases. But this should not lead to the conclusion that larger σ_g^2 reduces the precision of prediction because both criteria MD and MSE rely on the magnitude of σ_g^2. Unlike variance estimate of σ_g^2, the differences in both MD and MSE for BLUP of GCA or GV were distinguishable among different imbalance levels in this study. Moreover, the increase in MD or MSE was evident in unbalanced data sets, as compared with results from balanced data.

Table 2.1.8. Mean distance (MD) and mean square error (MSE) of GCA prediction in 3 models from analysis of simulated data sets (1000 sets for each parameter structure and imbalance level)

	MD of GCA prediction					MSE of GCA prediction				
PN*	MODEL1	MODEL1^A	MODEL2	MODEL2^A	MODEL3	MODEL1	MODEL1^A	MODEL2	MODEL2^A	MODEL3
1	0.7957	0.7327	0.7575	0.7323	0.7271	1.001	0.8445	0.9068	0.8429	0.8315
2	0.7460	0.6596	0.6961	0.6604	0.6566	0.8695	0.6793	0.7603	0.6808	0.6724
3	0.8631	0.803	0.831	0.8044	0.7984	1.1659	1.0046	1.0797	1.0088	0.9948
4	0.7932	0.7223	0.7522	0.7245	0.719	0.9885	0.8182	0.8928	0.8232	0.8096
5	1.0865	0.9864	1.0244	0.9852	0.9819	1.8483	1.5316	1.6432	1.5265	1.5147
6	0.9852	0.8510	0.8992	0.8519	0.8491	1.5169	1.1395	1.2773	1.1404	1.1328
7	1.1893	1.1086	1.1389	1.1094	1.101	2.2043	1.9167	2.0296	1.9208	1.8941
8	1.0974	0.9977	1.0347	0.9989	0.9923	1.9096	1.5917	1.7114	1.5949	1.5746
2a	0.7711	0.7019	0.7307	0.7036	0.6986	0.9357	0.7699	0.8397	0.775	0.7636
2b	0.7388	0.6629	0.6920	0.6638	0.6586	0.8592	0.6936	0.7599	0.6948	0.6847
2c	0.7975	0.7251	0.7558	0.7261	0.7210	1.0111	0.8344	0.9122	0.8371	0.8227
2d	0.7694	0.6910	0.7202	0.6911	0.6873	0.9325	0.7525	0.8215	0.7509	0.7424
6a	1.0317	0.9123	0.9607	0.9129	0.9089	1.6681	1.3104	1.4574	1.3126	1.3009
6b	1.0026	0.8884	0.9291	0.8887	0.8855	1.6002	1.245	1.3804	1.2479	1.2382
6c	1.0759	0.9510	1.0012	0.9532	0.9471	1.8089	1.4166	1.573	1.4245	1.407
6d	1.0074	0.8898	0.9353	0.8897	0.8863	1.6130	1.2529	1.3997	1.2542	1.2458
4D6	0.9858	0.7782	0.8199	0.7766	0.7752	3.0418	1.9037	2.1328	1.894	1.8867

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2
^{A} — diallel adjustment was applied to GCA prediction.

Table 2.1.9. Mean distance (MD) and mean square error (MSE) of GV prediction in 3 models from analysis of simulated data sets (1000 sets for each parameter structure and imbalance level)

	MD. of GV prediction					MSE of GV prediction				
PN	MODEL1	MODEL1 ^{A}	MODEL2	MODEL2 ^{A}	MODEL3	MODEL1	MODEL1 ^{A}	MODEL2	MODEL2 ^{A}	MODEL3
1	1.3388	1.1716	1.2345	1.1675	1.164	2.83	2.1622	2.411	2.1479	2.1338
2	1.2756	1.0535	1.1442	1.0532	1.0519	2.5598	1.7500	2.0794	1.7491	1.7433
3	1.4083	1.2237	1.2972	1.2203	1.2185	3.1159	2.3582	2.672	2.3472	2.3401
4	1.2980	1.0960	1.1721	1.0953	1.0942	2.6652	1.8807	2.1832	1.8794	1.8751
5	1.7952	1.5384	1.6283	1.5321	1.5294	5.0647	3.7062	4.1822	3.6782	3.6655
6	1.6908	1.3633	1.4754	1.3626	1.3619	4.5169	2.9244	3.4882	2.9216	2.9178
7	1.8793	1.6331	1.7214	1.6294	1.6256	5.5706	4.1793	4.6861	4.1611	1.1458
8	1.7674	1.4541	1.5648	1.4529	1.4521	4.8658	3.3513	3.8807	3.3459	3.3422
2a	1.2687	1.0711	1.1469	1.0705	1.069	2.1116	1.4971	1.7347	1.4964	1.4917
2b	1.2686	1.0753	1.1476	1.0744	1.0716	2.5276	1.8103	2.0792	1.8061	1.7989
2c	1.3207	1.1206	1.1988	1.1211	1.1176	2.3072	1.6444	1.9125	1.6449	1.6333
2d	1.3165	1.1205	1.1934	1.1186	1.1164	2.7455	1.9786	2.2624	1.9661	1.9586
6a	1.7186	1.3957	1.5196	1.3941	1.3924	3.8359	2.5157	3.0278	2.509	2.5016
6b	1.7143	1.4240	1.5265	1.4217	1.4193	4.6920	3.2115	3.7531	3.2064	3.1977
6c	1.7793	1.4463	1.5725	1.4452	1.4434	4.1175	2.7274	3.2566	2.7237	2.7176
6d	1.7583	1.4399	1.5633	1.4397	4.8180	3.2773	3.8835	3.2754	3.2754	3.2719
4D6	1.7078	1.1647	1.2792	1.1621	1.1618	9.1596	4.2655	5.2743	4.2459	4.2427

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

^{A} — diallel adjustment was applied to GCA prediction.

Correlation between true genetic value and predicted value

Two sets of correlations between true GCA and prediction were calculated (see table 2.1.10). Overall correlation was the correlation for 12000 parents from 1000 simulations per parameter structure and imbalance level. For each simulation run, correlation was also calculated for 12 parents of each simulated data set and then averaged over 1000 simulations to obtain mean correlation. Similar to mean distance and mean square error, after diallel effect adjustment, overall correlation or mean correlation improved consistently over all parameter structures and imbalance levels by about .04 in model 1 and about .02 in model 2.

Regardless of analytical model, correlation was high, as all models utilized BLUP methodology. For adjusted GCA prediction in Model 1 and Model 2 and GCA prediction in model 3, overall correlations exceeded 0.8 except for parameter structure 3 and 7, and mean correlations were all greater than 0.8. Mean correlation even reached .9 for parameter structure 6. Among the three models, Model 3 had the highest correlation for both overall correlation and mean correlation, although its advantage over the other two was slight. Between the other two models, Model 1 had a higher overall correlation and Model 2 had slight higher mean correlation. The differences in correlation among these models were so subtle that they were only distinguishable after 3rd decimal point.

Similar to the GCA correlation, both the overall correlation (sample size 30,000) and mean correlation (sample size 30) were calculated for true full-sib family genetic value (see Table 2.1.11). Both overall and mean correlations were also high. Comparisons among different models were virtually the same as GCA correlation.

The correlation between true genetic value and its prediction is probably the most important factor in a breeding program. It has direct implication to how much genetic gain will be achieved by selection. Based on results in this study, we can safely say that all three models give almost equally good predictions, with Model 3 being slightly better.

Table 2.1.10. Overall correlation and mean sample correlation between true parental GCA and prediction in 3 models from analysis of simulated data sets (1000 sets for each parameter structure and imbalance level)

	Overall Correlation					Mean sample Correlation				
PN*	MODEL1	MODEL2	MODEL1^{A}	MODEL2^{A}	MODEL3	MODEL1	MODEL2	MODEL1^{A}	MODEL2^{A}	MODEL3
1	0.7752	0.7984	0.8158	0.8149	0.8176	0.8119	0.8426	0.8481	0.8530	0.8595
2	0.8089	0.8351	0.8547	0.8540	0.8559	0.8416	0.8698	0.8863	0.8881	0.8923
3	0.7376	0.7593	0.7808	0.7781	0.7816	0.7694	0.7997	0.8091	0.8151	0.8268
4	0.7735	0.7980	0.8175	0.8157	0.8191	0.8063	0.8317	0.8443	0.8492	0.8567
5	0.7971	0.8217	0.8362	0.8360	0.8373	0.8272	0.8551	0.8649	0.8688	0.8718
6	0.8361	0.8641	0.8799	0.8796	0.8805	0.8708	0.9014	0.9140	0.9156	0.9169
7	0.7488	0.7709	0.7877	0.7856	0.7890	0.7837	0.8120	0.8185	0.8243	0.8343
8	0.7897	0.8136	0.8291	0.8280	0.8305	0.8251	0.8512	0.8618	0.8653	0.8725
2a	0.7888	0.8126	0.8307	0.8289	0.8317	0.8218	0.8497	0.8621	0.8656	0.8730
2b	0.8103	0.8343	0.8507	0.8499	0.8523	0.8381	0.8627	0.8772	0.8792	0.8855
2c	0.7704	0.7951	0.8161	0.8144	0.8180	0.8072	0.8399	0.8530	0.8569	0.8627
2d	0.7989	0.8252	0.8420	0.8418	0.8438	0.8329	0.8630	0.8740	0.8775	0.8827
6a	0.8170	0.8422	0.8599	0.8593	0.8607	0.8490	0.8779	0.8909	0.8933	0.8966
6b	0.8246	0.8509	0.8668	0.8663	0.8674	0.8592	0.8893	0.9010	0.9026	0.9047
6c	0.8041	0.8322	0.8507	0.8495	0.8515	0.8435	0.8770	0.8872	0.8896	0.8937
6d	0.8222	0.8478	0.8653	0.8649	0.8659	0.8573	0.8871	0.9012	0.9027	0.9045
4D6	0.8315	0.8853	0.8985	0.8988	0.8992	0.8471	0.9043	0.9155	0.9164	0.9169

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2
^{A} — diallel adjustment was applied to GCA prediction.

Table 2.1.11. Overall correlation and mean sample correlation between true full-sib family GV and prediction in 3 models from analysis of simulated data sets (1000 sets for each parameter structure and imbalance level)

	Overall Correlation					Mean sample Correlation				
PN*	MODEL1	MODEL2	MODEL1 ^{A}	MODEL2 ^{A}	MODEL3	MODEL1	MODEL2	MODEL1 ^{A}	MODEL2 ^{A}	MODEL3
1	0.7154	0.7638	0.7939	0.7939	0.7951	0.7798	0.8310	0.8527	0.8551	0.8571
2	0.7480	0.8016	0.8364	0.8362	0.8367	0.8121	0.8696	0.8992	0.8997	0.9008
3	0.7376	0.7803	0.8104	0.8105	0.8109	0.7926	0.8367	0.8603	0.8621	0.8633
4	0.7710	0.8174	0.8451	0.8450	0.8454	0.8272	0.8743	0.8966	0.8973	0.8979
5	0.7493	0.7985	0.8253	0.8258	0.8263	0.8110	0.8616	0.8806	0.8834	0.8846
6	0.7805	0.8358	0.8645	0.8644	0.8646	0.8486	0.9061	0.9298	0.9304	0.9307
7	0.7667	0.8082	0.8318	0.8320	0.8326	0.8256	0.8685	0.8881	0.8894	0.8906
8	0.8007	0.8449	0.8678	0.8679	0.8680	0.8654	0.9117	0.9329	0.9334	0.9334
2a	0.7464	0.9767	0.8287	0.8284	0.8289	0.8103	0.8642	0.8893	0.8901	0.8916
2b	0.7465	0.7973	0.8269	0.8269	0.8276	0.8023	0.8564	0.8803	0.8814	0.8829
2c	0.7214	0.7760	0.8120	0.8113	0.8127	0.7872	0.8468	0.8747	0.8759	0.8782
2d	0.7313	0.7851	0.8161	0.8168	0.8175	0.7951	0.8532	0.8771	0.8789	0.8802
6a	0.7752	0.8279	0.8595	0.8597	0.8601	0.8440	0.8990	0.9252	0.9262	0.9269
6b	0.7698	0.8213	0.8495	0.8495	0.8499	0.8391	0.8938	0.9164	0.9171	0.9177
6c	0.7621	0.8180	0.8503	0.8503	0.8506	0.8320	0.8891	0.9180	0.9186	0.9193
6d	0.7637	0.8152	0.8466	0.8466	0.8468	0.8323	0.8879	0.9140	0.9143	0.9147
4D6	0.7757	0.8786	0.9028	0.9030	0.9031	0.8098	0.9164	0.9355	0.9360	0.9361

Note: * — parameter structure and imbalance label, see table 2.1.1, 2.1.2

^{A} — diallel adjustment was applied to GCA prediction.

Conclusion

Model 3, which eliminates the diallel effect, is the best in both estimating variance component and predicting genetic values based on the simulated data from this study. Model 2 that treated the diallel as a random effect resulted in downward biased estimate of GCA variance component. Model 1, which treats diallel as a fixed effect produced unbiased GCA variance estimate. Both Model 1 and Model 2 suffered slight reduction in accuracy of GCA variance estimates. Between these 2 models, Model 1 appears to have less accuracy than Model 2 in terms of mean distance and MSE. But considering the downward bias of Model 2, the percentage reductions in accuracy for both models are virtually the same.

In both Model 1 and Model 2, BLUP of additive genetic values must be adjusted for diallel effect. The accuracy of adjusted prediction, measured as correlation between true genetic value and prediction, is very close to the best prediction in Model 3. The adjusted prediction from Model 2 has a slight edge over the adjusted one from Model 1.

Model 3 is preferred as long as the assumptions about the RCBD design and random selection of parents is correct, as in the case of NCSU-ICTIP diallel tests. Since model 1 performs well overall in both variance component estimation and BLUP, it can also be used in practical analysis. Model 2 is not recommended in most situations since its GCA variance estimation is biased and occasionally diallel variance could be unrealistically large.

The implications of this study are: 1) disconnected diallel designs can be efficiently analyzed using Model 3, where disconnected diallels are treated as a large incomplete diallel; 2) If a complete model needs to be used (e.g. if the assumption of random selection of parents is invalid), treating diallel as fixed effect as in Model 1 is preferred in 2-disconnected diallel setting; 3) treating diallel as random effect will result in downward biased GCA variance estimate, thus underestimate additive genetic variance and heritability.

II. Analytical methods for combining different disconnected diallel test series within a region

Abstract

Statistical approaches were evaluated for combining multiple disconnected diallel test series within a given region. The best GCA sample variance prediction in the class of linear combination of local variance estimates was derived in this study. The efficiency of a common checklot adjustment and GCA variance adjustment were investigated using simulated data sets. Results from simulated data showed that analysis of disconnected diallel test series with checklot adjustment is very critical to improve the prediction of genetic values obtained using BLUP analysis. Checklot adjustment increased correlation more than .05 in GCA prediction, and almost .1 in full-sib GV in all parameter settings. Additional adjustment with improved GCA sample variance prediction could improve the correlation slightly beyond checklot adjustment. But this limited improvement is due to the fact that the accuracy of checklot-adjusted prediction alone is already close to the theoretical limit. For practical reasons, the checklot adjustment method may be the best way for predicting breeding value for combined multi-disconnected test series for a breeding region.

Key words: Disconnected Diallel Mating; BLUP analysis; GCA; Breeding Value; Checklot.

Introduction

In a typical breeding program, it is necessary to rank and select the best parents and crosses from several disconnected test series so that an effective breeding strategy can be developed based on the whole breeding population within a test region. If the same set of families are tested or there are common families across test series, direct prediction of GCA and breeding values from single test series analysis are comparable across test series or can be easily adjusted for the difference. In the case of disconnected mating designs, different sets of parents are often included in each test series and thus different sets of families are tested in different test series. If the genetic by environmental effect within a test region is small as in loblolly pine (chapter 3; Li and McKeand 1989, McKeand et al. 1990), separate test series analysis does not utilize homogeneity condition of genetic variance, and hence may not be the most efficient way to estimate GCA variance and to predict genetic values. In addition, the environmental conditions such as field quality and site uniformity, rust infection level, and test management could impact the traits that are of interested. Hence we would expect that the breeding value prediction of the same family might vary from one test series to the other.

As a popular mating design used in crop and tree breeding programs (Li et al., 1997; Li et al, 1996; Yanchuk, 1996; Huber, 1992), disconnected diallel design has several advantages over other designs with regards to the goals of genetic tests (van Buijtenen, 1990; Zobel et al. 1984). But it is difficult to analyze especially when dealing with disconnected diallel in different test series. The new mixed procedure method was developed to overcome the difficulties in analyzing diallel tests (chapter 1). With the appropriate linear model (chapter 2), disconnected diallel tests for each test series can be efficiently analyzed.

BLUP analysis has been accepted as a preferred methodology over other procedures such as OLS or GLS in tree breeding programs and has been used in analyzing diallel tests (Borralho, 1995; Huber 1993; White & Hodge, 1989). BLUP can easily be done for each test series using SAS PROC MIXED or GAREML. However, like any disconnected design, the same concern on mean variation due to test conditions raises an important question regarding data analysis: how do we rank parents and families from different disconnected diallels and use the existing BLUPs from a single test series analysis to predict GCA and breeding value for the entire region? Is there a way to improve GCA variance estimate for the purpose of breeding value prediction? How do we adjust for the environmental variation among test series? A commonly used method for adjusting this variation in forest tree breeding program is to use a common checklot for planting in different test series. Usually, an unimproved checklot is included in testing to connect the test series and get a rough estimate of genetic gain. Can the common checklot be used for combined analysis of disconnected diallels in different test series? What are the major limitations by using this method? The goal of this research is to evaluate several analytical approaches for improving the accuracy of prediction. The objective of the study is to:

1. Derive a method to improve GCA variance component estimates based on estimates from single test series analysis and show how to use the improved variances estimate to improve breeding value prediction;

2. Evaluate the checklot adjustment and develop methods to use the checklot mean to adjust breeding value prediction for among test series variation;

3. Use simulated data to evaluate the use of both GCA variance adjustment and checklot adjustment for improving GCA and breeding value prediction.

Theoretical consideration

Best variance component prediction

Consider a scenario in which there are e test sites, i.e. test series over the entire test region, each with the same mating design and field design. Regardless of full-sib or half-sib design, we assume each test series has n different parents. Within each test series, the GCA variance component and GCAs for each parent can be estimated or predicted using mixed model analysis with SAS. Denote genetic value of *j*th parent of *i*th test series as g_{ij}, i=1,…,e; j=1,…,n. Assume that g_{ij}'s are independent identically distributed as N(0, σ_g^2), n·e parents are from a population with GCA variance σ_g^2 and {g_{ij}} is a random sample from N(0, σ_g^2). For a particular test series i, {g_{i1}, g_{i2}, …, g_{in}} is a random sub sample of size n. For a set of given g_{ij}'s, we can define a "statistic" called true sample variance S^2_{gi} (it is in quote because g_{ij} is not observed in the real data) for each test series as

$$S_{gi}^2 = \frac{\sum_j (g_{ij} - \overline{g_i})^2}{n-1} \qquad (1)$$

where $\overline{g_i} = \frac{\sum_j g_{ij}}{n}$

As a property of sampling from normal distribution, we have,

$$E(S^2_{gi}) = \sigma_g^2 \quad \text{and} \quad \frac{(n-1)S_{gi}^2}{\sigma_g^2} \sim \chi^2_{n-1}$$

Since the variance of this Chi-squared distribution is 2(n-1), we find the variance of S_i^2

$$\mathrm{Var}(S_{gi}^2) = \frac{2}{n-1}(\sigma_g^2)^2 \qquad (2)$$

In the separate analysis of each test series, we can obtain REML estimate $\hat{\sigma}_{gi}^2$ (or other good estimate) of GCA variance using mixed model analysis. REML estimate is usually considered to be the best or nearly best (Huber, 1993; Searle, 1992; Swallow et al., 1984). So let's assume that in a single *i*th test series, $\hat{\sigma}^2_{gi}$ is the "best" predictor of S_i^2 (called predictor because S_i^2 itself is a random variable). But when we consider the whole test region and try to combine the results from many test series, is $\hat{\sigma}^2_{gi}$ still the best predictor of S_i^2? Or does $\hat{\sigma}^2_{gi'}$ from any other test series give any information to estimate S_i^2?

Let's focus on the class of predictors using the linear combination of GCA variance estimates from all test series that is $W(\hat{\boldsymbol{\sigma}}^2_g) = \sum_{i=1}^{e} a_i \hat{\sigma}^2_{gi}$, where $\hat{\boldsymbol{\sigma}}^2_g = (\hat{\sigma}^2_{g1}, ..., \hat{\sigma}^2_{ge})$ $0 \le a_i \le 1$ and $\sum_{i=1}^{e} a_i = 1$.

Since,

$$E\hat{\sigma}^2_{gi} = \sigma_g^2, \qquad E\, S^2_{gi} = \sigma_g^2 \qquad \text{and} \qquad E[\hat{\sigma}^2_{gi} | S^2_{gi}] = S_{gi}^2$$

it is easy to verify that $E(W- S_i^2)=0$, so W is an unbiased predictor of S_i^2. For simplicity, we also assume that all test series share the same variance of GCA variance estimate. In other words, all test series have the same quality in terms of estimating GCA variance. Under this assumption, we can argue that all test series other than *i*th test series contribute the same amount of information, if any, to prediction. In the best prediction, the weight a_j should be equal for all $\hat{\sigma}^2_{gj}$, i.e. $a_j = c$ for all $j \neq i$. Hence we only need to consider the following subset class of predictors:

$$\begin{aligned} W(\hat{\boldsymbol{\sigma}}^2_g) &= a_i \hat{\sigma}^2_{gi} + c\sum_{j \neq i} \hat{\sigma}^2_{gj} \\ &= (a_i - c)\hat{\sigma}^2_{gi} + c\sum_{i=1}^{e} \hat{\sigma}^2_{gi} \\ &= (a_i - c)\hat{\sigma}^2_{gi} + ec\overline{\hat{\sigma}^2_{g.}} \qquad \text{where } \overline{\hat{\sigma}^2_{g.}} \text{ is defined as } \frac{\sum_{i=1}^{e} \hat{\sigma}^2_{gi}}{e} \end{aligned}$$

Let λ=a$_i$ – c. Since $\sum_{i=1}^{e} a_i = 1$ we have a$_i$ + c(e-1)=1. Hence ec=1-λ, i.e. W can be written as:

$$W(\lambda, \hat{\boldsymbol{\sigma}}_g^2) = \lambda\hat{\sigma}_{gi}^2 + (1-\lambda)\overline{\hat{\sigma}_{g.}^2} \qquad (3)$$

We wish to find λ that minimizes its mean square error, i.e.

$$\begin{aligned}
E(W - S_{gi}^2)^2 &= Var(W - S_{gi}^2) \\
&= E\{Var[(W - S_{gi}^2) \mid S_{gi}^2]\} + Var\{E[(W - S_{gi}^2) \mid S_{gi}^2]\} \\
&= E[Var(W \mid S_{gi}^2)] + Var[E(W \mid S_{gi}^2) - S_{gi}^2]
\end{aligned}$$

$$\begin{aligned}
E[Var(W \mid S_{gi}^2)] &= E\{Var[\lambda\hat{\sigma}_{gi}^2 + (1-\lambda)\overline{\hat{\sigma}_{g.}^2} \mid S_{gi}^2]\} \\
&= E\left\{Var\left[\left(\lambda + \frac{1-\lambda}{e}\right)\hat{\sigma}_{gi}^2 + \frac{1-\lambda}{e}\sum_{j\neq i}^{e}\hat{\sigma}_{gj}^2 \middle| S_{gi}^2\right]\right\} \\
&= E\left\{\left(\lambda + \frac{1-\lambda}{e}\right)^2 Var(\hat{\sigma}_{gi}^2 \mid S_{gi}^2) + \left(\frac{1-\lambda}{e}\right)^2 \mathrm{var}\left(\sum_{j\neq i}^{e}\hat{\sigma}_{gj}^2 \middle| S_{gi}^2\right)\right\} \qquad (\hat{\sigma}_{gi}^2, \hat{\sigma}_{gj}^2 \textit{ are independent conditional on } S_{gi}^2) \\
&= \left(\lambda + \frac{1-\lambda}{e}\right)^2 E[Var(\hat{\sigma}_{gi}^2 \mid S_{gi}^2)] + \left(\frac{1-\lambda}{e}\right)^2 (e-1)E[\mathrm{var}(\hat{\sigma}_{gj}^2)] \qquad (S_{gi}^2, \hat{\sigma}_{gj}^2 \textit{ are independent}) \\
&= \left(\lambda + \frac{1-\lambda}{e}\right)^2 \{Var(\hat{\sigma}_{gi}^2) - Var[E(\hat{\sigma}_{gi}^2 \mid S_{gi}^2)]\} + \left(\frac{1-\lambda}{e}\right)^2 (e-1)\,\mathrm{var}(\hat{\sigma}_{gj}^2) \\
&= \left(\lambda + \frac{1-\lambda}{e}\right)^2 \{Var(\hat{\sigma}_{gi}^2) - Var(S_{gi}^2)\} + \left(\frac{1-\lambda}{e}\right)^2 (e-1)\,\mathrm{var}(\hat{\sigma}_{gi}^2) \qquad (Var(\hat{\sigma}_{gi}^2) = Var(\hat{\sigma}_{gj}^2))
\end{aligned}$$

$$\begin{aligned}
E(W \mid S_{gi}^2) &= E\left[\lambda\hat{\sigma}_{gi}^2 + (1-\lambda)\overline{\hat{\sigma}_{g.}^2} \mid S_{gi}^2\right] \\
&= \lambda S_{gi}^2 + \frac{1-\lambda}{e} E\left[\hat{\sigma}_{gi}^2 + \sum_{j\neq i}^{e}\hat{\sigma}_{gj}^2 \mid S_{gi}^2\right] \\
&= \lambda S_{gi}^2 + \frac{1-\lambda}{e} S_{gi}^2 + \frac{1-\lambda}{e}(e-1)\sigma_g^2 \\
&= \left(\lambda + \frac{1-\lambda}{e}\right) S_{gi}^2 + \frac{1-\lambda}{e}(e-1)\sigma_g^2
\end{aligned}$$

$$\mathrm{Var}\left[\mathrm{E}(\mathrm{w} \mid \mathrm{S}_{\mathrm{gi}}^2) - \mathrm{S}_{\mathrm{gi}}^2]\right] = \mathrm{Var}\left[\left(\lambda + \frac{1-\lambda}{\mathrm{e}}\right)\mathrm{S}_{\mathrm{gi}}^2 + \frac{1-\lambda}{\mathrm{e}}(\mathrm{e}-1)\sigma_{\mathrm{g}}^2 - \mathrm{S}_{\mathrm{gi}}^2\right]$$
$$= \mathrm{Var}\left[\left(\lambda - 1 + \frac{1-\lambda}{\mathrm{e}}\right)\mathrm{S}_{\mathrm{gi}}^2\right]$$
$$= (1-\lambda)^2\left(\frac{\mathrm{e}-1}{\mathrm{e}}\right)^2 \mathrm{Var}(\mathrm{S}_{\mathrm{gi}}^2)$$

Hence,

$$\mathrm{E}(\mathrm{W} - \mathrm{S}_{\mathrm{gi}}^2)^2 = \left(\lambda + \frac{1-\lambda}{\mathrm{e}}\right)^2 \left\{\mathrm{Var}(\hat{\sigma}_{\mathrm{gi}}^2) - \mathrm{Var}(\mathrm{S}_{\mathrm{gi}}^2)\right\} + \left(\frac{1-\lambda}{\mathrm{e}}\right)^2 (\mathrm{e}-1)\,\mathrm{var}(\hat{\sigma}_{\mathrm{gi}}^2)$$
$$+ (1-\lambda)^2\left(\frac{\mathrm{e}-1}{\mathrm{e}}\right)^2 \mathrm{Var}(\mathrm{S}_{\mathrm{gi}}^2) \qquad (4)$$

Let a=$Var(\hat{\sigma}_{gi}^2)$, b=$Var(S_{gi}^2)$, i=1,…,e, rewrite the above equation,

$$\mathrm{E}(\mathrm{W} - \mathrm{S}_{\mathrm{gi}}^2)^2 = \left(\lambda + \frac{1-\lambda}{\mathrm{e}}\right)^2 \{\mathrm{a} - \mathrm{b}\} + \left(\frac{1-\lambda}{\mathrm{e}}\right)^2 (\mathrm{e}-1)\mathrm{a}$$
$$+ (1-\lambda)^2\left(\frac{\mathrm{e}-1}{\mathrm{e}}\right)^2 \mathrm{b}$$
$$= \mathrm{a} - \mathrm{b} + (1-\lambda)^2 \frac{(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}} - 2(1-\lambda)\frac{(\mathrm{e}-1)}{\mathrm{e}}(\mathrm{a}-\mathrm{b})$$

Set the first derivative to zero,

$$\frac{\mathrm{dE}(\mathrm{W} - \mathrm{S}_{\mathrm{gi}}^2)^2}{\mathrm{d}\lambda} = -2(1-\lambda)\frac{(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}} + 2\frac{(\mathrm{e}-1)}{\mathrm{e}}(\mathrm{a}-\mathrm{b}) = 0$$

Which leads to the solution,

$$\lambda_{\min} = \frac{\mathrm{b}}{\mathrm{a}} \qquad (5)$$

Disregarding the value of λ, the second derivative, $\frac{\mathrm{d}^2\mathrm{E}(\mathrm{W} - \mathrm{S}_{\mathrm{gi}}^2)^2}{\mathrm{d}\lambda^2} = \frac{2(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}}$ is non-negative, it is a minimum.

How well does it compare with $\hat{\sigma}_{gi}^2$ or $\overline{\hat{\sigma}_{g.}^2}$, i.e. when λ=1 or 0? We can look at the their difference in mean square errors,

$$\begin{aligned}
&\mathrm{MSE}[\mathrm{W}(1,\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] - \mathrm{MSE}[\mathrm{W}(\lambda_{\min},\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] \\
&= \mathrm{a} - \mathrm{b} - \left[\mathrm{a} - \mathrm{b} + (1-\lambda_{\min})^2\frac{(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}} - 2(1-\lambda_{\min})\frac{(\mathrm{e}-1)}{\mathrm{e}}(\mathrm{a}-\mathrm{b})\right] \\
&= \frac{(\mathrm{a}-\mathrm{b})^2}{\mathrm{a}}\cdot\frac{\mathrm{e}-1}{\mathrm{e}}
\end{aligned}$$

$$\begin{aligned}
&\mathrm{MSE}[\mathrm{W}(0,\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] - \mathrm{MSE}[\mathrm{W}(\lambda_{\min},\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] \\
&= \mathrm{a} - \mathrm{b} + \frac{(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}} - 2\frac{(\mathrm{e}-1)}{\mathrm{e}}(\mathrm{a}-\mathrm{b}) - \left[\mathrm{a} - \mathrm{b} + (1-\lambda_{\min})^2\frac{(\mathrm{e}-1)\mathrm{a}}{\mathrm{e}} - 2(1-\lambda_{\min})\frac{(\mathrm{e}-1)}{\mathrm{e}}(\mathrm{a}-\mathrm{b})\right] \\
&= \frac{\mathrm{b}^2(\mathrm{e}-1)}{\mathrm{ae}}
\end{aligned}$$

The above two quantities are always positive. Also notice

$$\mathrm{EMS}[\mathrm{W}(0,\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] - \mathrm{EMS}[\mathrm{W}(1,\hat{\boldsymbol{\sigma}}_{\mathrm{g}}^2)] = \frac{(\mathrm{e}-1)}{\mathrm{e}}(2\mathrm{b}-\mathrm{a})$$

It indicates that whether $\overline{\hat{\sigma}_{g.}^2}$ is better than $\hat{\sigma}_{gi}^2$ depends on the sign of 2b-a. We may write $\mathrm{W}(\lambda_{\min},\hat{\sigma}_{\mathrm{g}}^2)$ as $\hat{\mathrm{s}}_{\mathrm{gi}}^2$ the best predictor in the class of linear combination of $\hat{\sigma}_{gi}^2$'s. It should be pointed out that both a and b are unknown parameters and need to be estimated. Standard error associated with $\hat{\sigma}_{gi}^2$ usually is available from variance components estimation procedure and can be used to estimate a. The following is one way to obtain the estimate of b:

$$\hat{b} = \frac{2}{n-1}(\overline{\hat{\sigma}_{g.}^2})^2 \qquad (6)$$

Variance component adjustment

Within each test series, accuracy of GCA variance component estimate has virtually no effect on ranking parents or full-sib families and selection. Increase in GCA variance will cause more variation in predicted GCAs but will not change the ranking of predicted values. In other words, the predicted genetic gain, which is directly related to GCA variance estimate, may change, but the actual gain remains the same. However with multiple test series, the absolute value of GCA variance does matter. This point is illustrated by the following graph, in which parent A and B are assumed to have the exact true GCA values. However, due to separate testing, two different variance estimates lead to two different sets of GCA predictions.

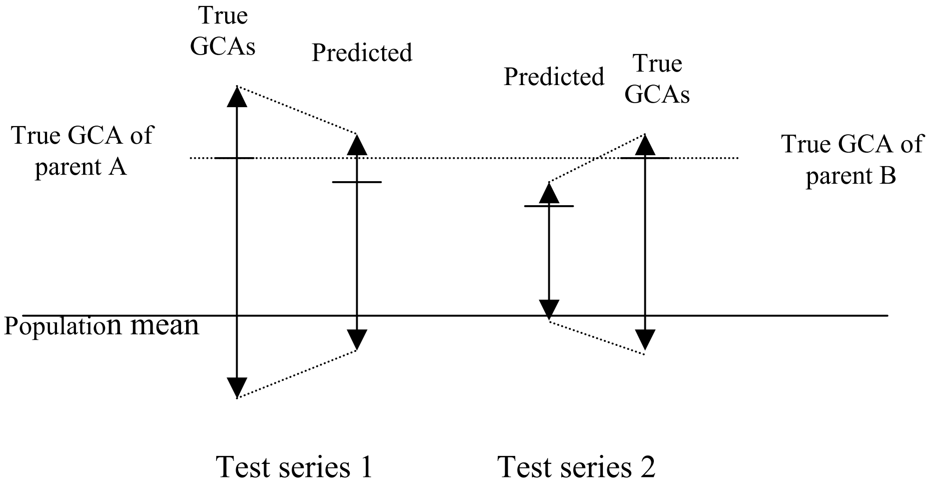


Fig 2.2.1 Different GCA predictions of parent A and B in two test series

With the “best” or better GCA sample variance prediction $\hat{s}_{gi}^2$, there are several ways to adjust the random genetic effect prediction. One approach is to put this “best” GCA sample variance together with estimated variance components for random effect into the mixed model and get the BLUP solution. In SAS PROC MIXED, this can be done by specifying variance components in the PARMS statement. Another easier and straightforward way is using the following adjustment on each GCA or full-sib genetic value prediction of ith test series:

$$\hat{G}^* = \hat{G}\sqrt{\frac{\hat{s}^2_{gi}}{\hat{\sigma}^2_{gi}}} \qquad (7)$$

Checklot adjustment justification

The variation of true parental mean GCA or breeding value is not negligible. In case of 12 parents, variance of the mean GCA equals to 1/12 σ_g^2. Thus when comparing two sets of 12 parents, the difference between two mean GCA's has variance of 1/6 σ_g^2. On the other hand, the shrinkage effect of random effect prediction leads to the predicted GCA's with a variance of considerably less than σ_g^2, which makes this mean variation even large. As a result, considerable inaccuracy will occur in the comparison of parents and crosses from different test series. We may see this more clearly in the following numerical example:

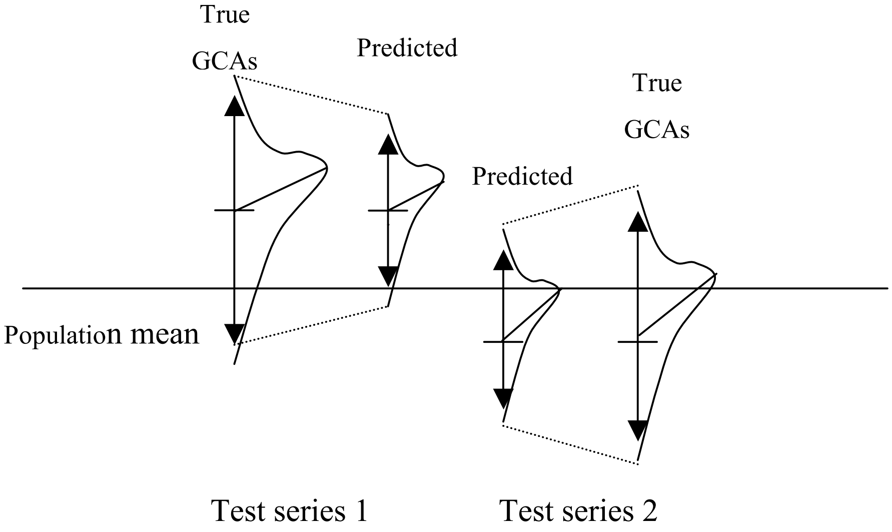


Figure 2.2.2 Two sets of parental GCAs from two test series

Suppose the correlation between predicted GCA and true GCA of 12 parents is .8 or $\sigma^2_{\hat{g}}/\sigma_g^2$=.64, which is considered as typical in practice. The probability that the difference between 2 mean GCAs is greater than $1.28\sqrt{\sigma^2_g/6}$ i.e. $.5226\sigma_g$ or $.6533\sigma_{\hat{g}}$ is .20. This discrepancy is further illustrated in the above diagram (Fig 2.2.2).

One alternative is to use grand mean of test series to adjust for difference in disconnected test series. However, the variations among tests and blocks often make the test series grant mean vary from test series to test series. The grand mean is largely determined by environmental factors. In the real data, the variance of test and block is huge (~80% of phenotypic variance for test, and about 10% for block). Thus, we cannot assume that the same set of parents will perform uniformly across test series. That means that the test series grand mean is not a suitable candidate for adjustment.

If in each test series, the test design includes several checklots as in the case of NCSU-TIP. If checklot trees are tested in each block within each test, they are exposed to the same environmental condition as other progenies. With diverse genetic background and large progeny number (2 plots in each replication), their stable performance is expected. Since the same checklots are used in all test series, their mean can be used as the standard with which all tested progenies can be compared. GCA and BV can be adjusted using the following:

GCA*=GCA+.5*(Diallel mean-Checklot mean)

BV*=BV+(diallel mean-checklot mean)

The adjustment is expected to improve the accuracy of genetic value prediction.

Method

Generation of simulated diallel data sets and checklots

For each simulated diallel test series, two disconnected 6-parent half diallel were tested with a randomized complete block field design with 4 tests, 6 blocks and 6 trees. Due to the limited resources, plot means are used in simulation. It can greatly reduce computational time while providing virtually the same amount of information as simulation based on individual observations. Imbalance due to missing plots and crosses can also be considered in plot mean simulation. The linear model for plot means follows equation (2) in the first section of this chapter.

Selection of parameters and determination of true variance values follow the same procedure as in the first section of this chapter (Lu, 1999; Huber, 1993). Three important parameters were also considered: 1) Narrow sense heritability, h^2; 2) dominance to additive genetic variance ratio, γ; 3) type B genetic correlation, r_B, each with two levels. By setting total phenotypic variation as 100 and assuming σ^2_p accounts for 10% of the rest of the variation, we obtained values for all variance components (listed in table 2.2.1). The plot mean variance σ^2_{pt} is σ^2_p+(1/6)σ^2_e.

The number of test series in a test region e in the real data varies from as small as 4 to more than 10. This number also subjectively depends on how we define the geographic regions and our breeding strategy. To reflect this range of variation in this study, two levels of e are considered: 4 and 8.

Table 2.2.1. Parameter structure for generating simulated data sets

Label	e*	h^2	γ	r_B	σ^2_g	σ^2_s	σ^2_{gt}	σ^2_{st}	σ^2_p	σ^2_e	σ^2_{pt}
1	4	.1	.3	.6	2.500	.750	1.667	.500	9.042	81.374	22.604
2	4	.1	.3	.8	2.500	.750	.625	.188	9.281	83.531	23.203
3	4	.1	.7	.6	2.500	1.750	1.667	1.167	8.875	79.874	22.187
4	4	.1	.7	.8	2.500	1.750	.625	.438	9.156	82.406	22.890
5	4	.2	.3	.6	5.000	1.500	3.333	1.00	8.083	72.751	20.208
6	4	.2	.3	.8	5.000	1.500	1.250	.375	8.563	77.062	21.407
7	4	.2	.7	.6	5.000	3.500	3.333	2.333	7.750	69.751	19.375
8	4	.2	.7	.8	5.000	3.500	1.250	.875	8.313	74.812	20.782
1	8	.1	.3	.6	2.500	.750	1.667	.500	9.042	81.374	22.604
2	8	.1	.3	.8	2.500	.750	.625	.188	9.281	83.531	23.203
3	8	.1	.7	.6	2.500	1.750	1.667	1.167	8.875	79.874	22.187
4	8	.1	.7	.8	2.500	1.750	.625	.438	9.156	82.406	22.890
5	8	.2	.3	.6	5.000	1.500	3.333	1.00	8.083	72.751	20.208
6	8	.2	.3	.8	5.000	1.500	1.250	.375	8.563	77.062	21.407
7	8	.2	.7	.6	5.000	3.500	3.333	2.333	7.750	69.751	19.375
8	8	.2	.7	.8	5.000	3.500	1.250	.875	8.313	74.812	20.782

* e= the number of test series in a test region

Checklot trees are simulated as the sum of test effect, block effect and a random phenotypic deviation. A total of 288 trees are generated for each checklot. For details in generating simulated diallel data sets, refer to the first section of this chapter.

Mixed model analysis of combined test series

Using the PROC MIXED approach in SAS (described in chapter 1), we performed BLUP analysis using SAS on the simulated diallel test data (SAS, 1996; Littell, R.C. et al., 1996). In the linear model (as the following), test and block are analyzed as fixed effect, although they are generated as random effects.

$$\overline{Y}_{ijokl.} = \mu + T_i + B_{j(i)} + D_o + G_{k(o)} + G_{l(o)} + S_{kl(o)} + TG_{ik(o)} + TG_{il(o)} + TS_{ikl(o)} + PT_{ijokl}$$

where, $\overline{Y}_{ijokl.}$ is the plot mean observation of the jth block within ith test for the klth cross of oth diallel;

μ is the overall mean;

T_i is the fixed effect of ith test, i=1 to t;

$B_{j(i)}$ is the fixed effect of jth block within ith test, j=1 to b;

D_o is the oth diallel effect, o=1 to d ~ NID(0, σ^2_d);

$G_{k(o)}$ or $G_{l(o)}$ is the GCA effect of kth female or lth male of oth diallel (k,l=1, … , p; k<l) ~ NID(0, σ^2_{GCA});

$S_{kl(o)}$ is the SCA effect of kth and lth parents of oth diallel ~ NID(0, σ^2_{SCA});

TG_{iok} or TG_{il} is the ith test by kth female or lth male GCA of oth diallel interaction ~ NID(0, $\sigma^2_{TEST*GCA}$);

TS_{iokl} is the ith test by of kth and lth parents SCA of oth diallel interaction ~ NID(0, $\sigma^2_{TEST*SCA}$);

PT_{ijokl} is the random plot mean error for the klth cross of oth diallel in the jth block within ith test ~ NID(0, σ^2_{PT}).

Diallel was also treated as fixed. The previous simulation study in the first part of this chapter showed that though the true diallel effect is null the analysis still produces a non-zero estimate on average. But this has little effect on the GCA variance estimate if the diallel effect is treated as fixed. By adjusting for diallel effect, the BLUP of GCA is also comparable to other models.

The GCA and full-sib genetic values (GV) from each single test series were further summarized using the methods that were discussed in the variance component and checklot adjustment section. $Var(\hat{\sigma}^2_{gi})$ was estimated by taking the average of variances of GCA variance estimates for each simulated test region and $Var(S^2_{gi})$ was determined by formula (6). These two estimates were used to calculate λ_{min} (listed in Appendix 3). Besides using the best sample variance predictor, several other GCA variance predictors were also used for adjustment to compare with the best predictor. These predictors include global GCA variance (where λ=0), the mean of global and individual test series GCA variance (λ=.5) and individual test series GCA variance (λ=1), i.e. no variance adjustment. Adjustment was done by replacing $\hat{s}^2_{gi}$ with these variance predictors in the formula (7). In addition, the true sample variance was used for adjustment to see what is the maximum improvement the adjustment can achieve.

Comparisons were made among these prediction values (GCA or GV) from different adjustment methods. Criteria include mean distance, mean square error, and correlation between predicted value and true genetic value.

Results and discussion

Mean distance and Mean square error

GCA

Mean distance of predicted GCA ranged from .6 to .8 for simulated diallel data sets with heritability of .1 and from .8 to 1.2 for data sets with heritability of .2 (Fig 2.2.3). Mean square error of GCA prediction ranged from .6 to 1.1 for data sets of heritability .1 and from .9 to 2.2 for those of heritability .2 (Fig 2.2.4). The number of test series within a test region appeared to have no effect on mean distance and mean square error.

For each parameter structure, the size of both mean distance and mean square error generally had the following order (from largest to smallest): non-adjusted for anything (non-adj), checklot adjusted (c-adj), checklot plus global variance adjusted (c-adj&global), checklot plus average variance adjusted (c-adj&vg1), checklot plus best variance adjusted (c-adj&vgb) and checklot plus true sample variance adjusted (c-adj&vgs).

The effect of checklot adjustment was evident (Fig 2.2.3, 2.2.4). Checklot adjustment alone significantly reduced both mean distance and mean square error. Although mean distance and MSE themselves varied from different parameter settings, the reduction after adjustment is consistent across parameter settings. The reduction was more evident as heritability increases, type B genetic correlation increases or non-additive genetic control decreases.

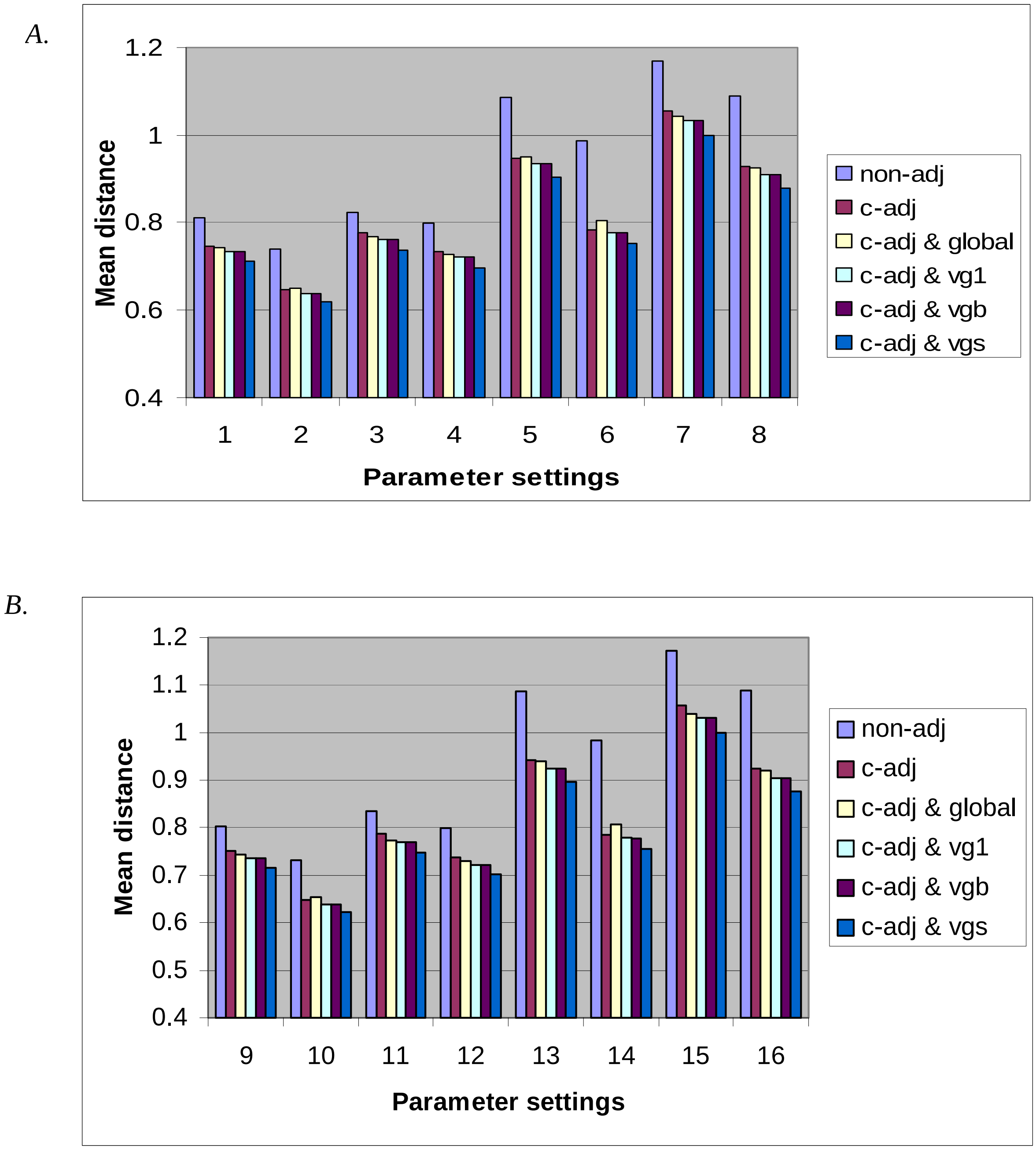


Figure 2.2.3 Mean distance (MSE) of parental GCAs from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 (A) test series and 8 (B) test series

Note: prediction are non-adjusted (non-adj), checklot adjusted (c-adj) with or without variance adjustment. Variance adjustment uses global variance (global), mean of global and individual variance (vg1), best variance predictor (vgb) or true sample variance (vgs)

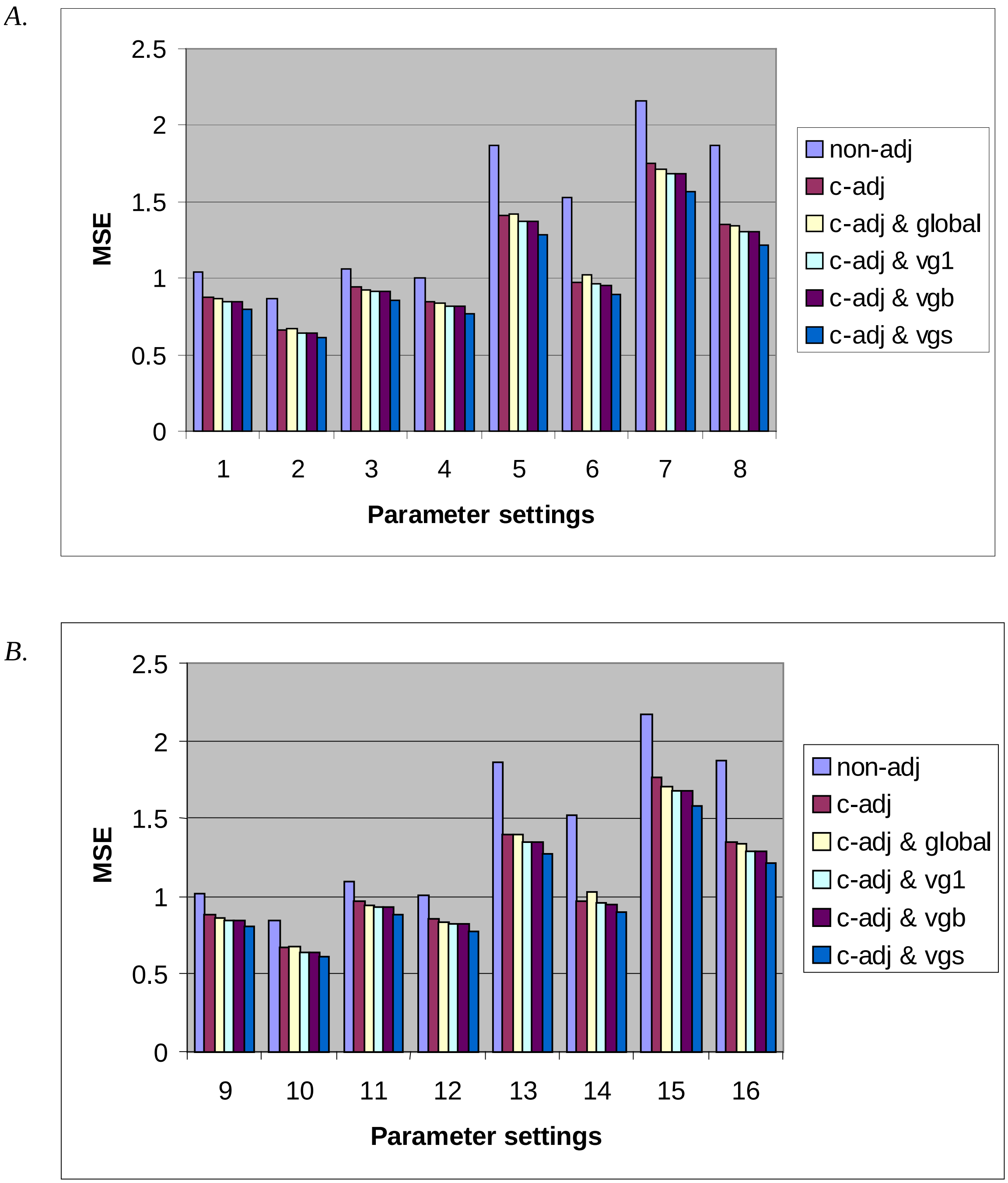


Figure 2.2.4 Mean square error (MSE) of parental GCAs from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 test series (a) and 8 test series (B)

Note: notation is the same as figure 2.2.3

Further adjustment for GCA variance gave less benefit to improvement of GCA prediction. In fact, even using the true GCA sample variance, which is the theoretical limit for adjustment, gives little reduction in both mean distance and MSE. However, if maximum accuracy of prediction is desired, the best variance predictor, vgb, can produce best GCA prediction from a practical point of view because it had consistently less or equal mean distance and MSE than other methods for all parameter settings. Using global GCA variance across the region produced MD or MSE higher than checklot adjustment alone in parameter setting 2, 6, 10, 14. A simple adjustment using the average of global variance and local variance surprisingly achieved GCA prediction of almost the same accuracy as the best predictor vgb.

Full-sib family GV

Mean distance and mean square error of full-sib family genetic value are illustrated in Fig 2.2.5 and 2.2.6. Mean distance of predicted full-sib GV ranged from 1.0 to 1.4 for simulated diallel data sets with heritability of .1 and from 1.1 to 1.8 for data sets with heritability of .2. Mean square error of GV prediction ranged from 1.6 to 3.0 for data sets of heritability .1 and from 2.1 to 5.5 for those of heritability .2. Similar to GCA prediction, both mean distance and mean square error did not differ much for both levels of test series number within a test region.

For each parameter structure, the size of both mean distance and mean square error followed the same order as GCA prediction, from largest to smallest: non-adj, c-adj, c-adj&global, c-adj&vg1, c-adj&vgb and c-adj&vgs.

Full-sib family GV measures the total genetic value of full-sib crosses, which is appropriate for calculating genetic gains from mass production of full-sib crosses. The above graphs show similar results as GCA predictions. Since it includes both GCA and SCA predictions and there is a slight negative correlation between these two, the relative order of predictions from different approaches slightly changes from that of GCA prediction alone. Most noticeably, the result from the checklot adjustment alone is comparable to the best one and is close to the theoretical limit. This implies that there is little room for GV prediction improvement beyond checklot adjustment.

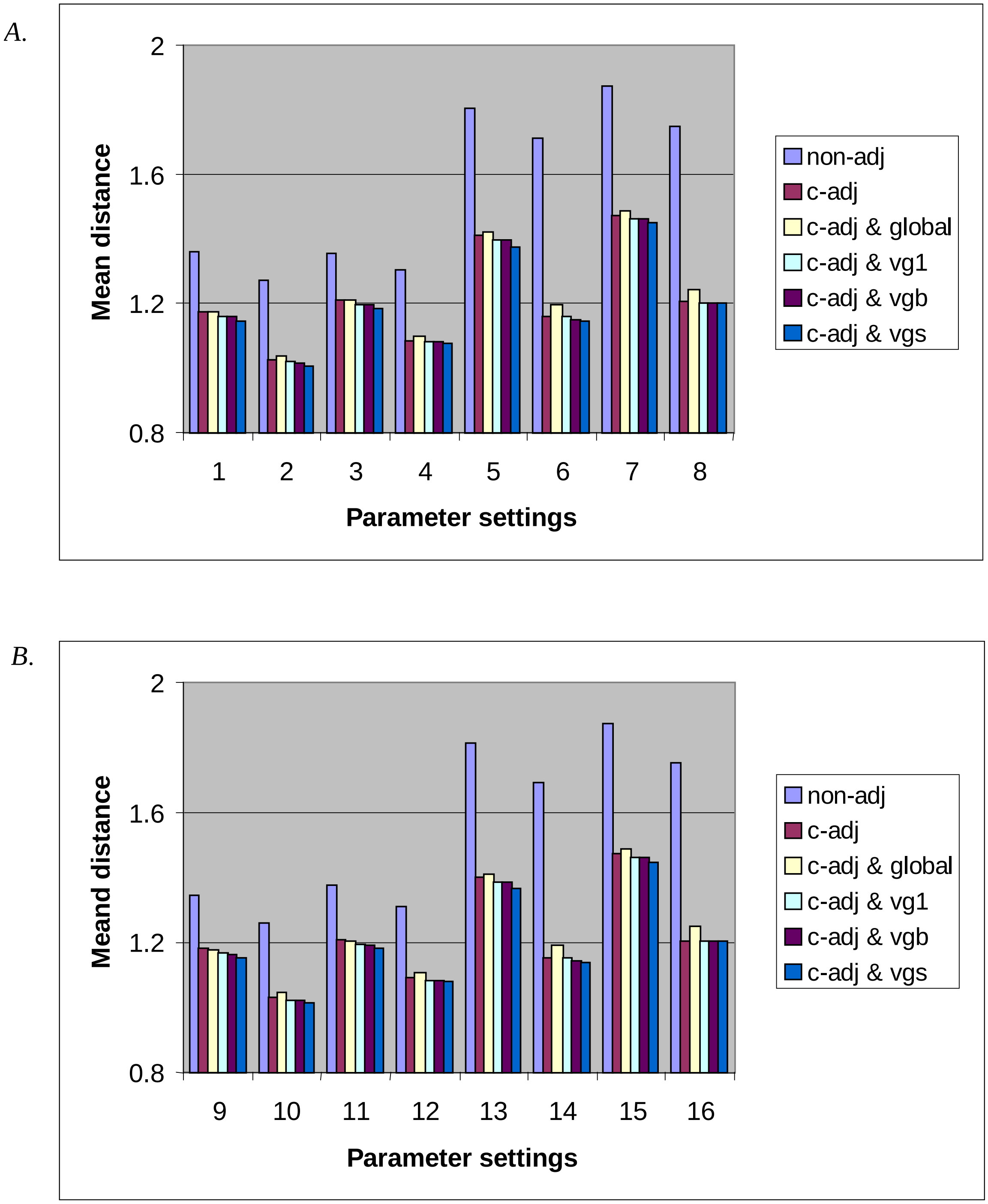


Figure 2.2.5 Mean distance (MSE) of full-sib genetic values (GV) from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 (A) test series and 8 (B) test series

Note: notation is the same as figure 2.2.3

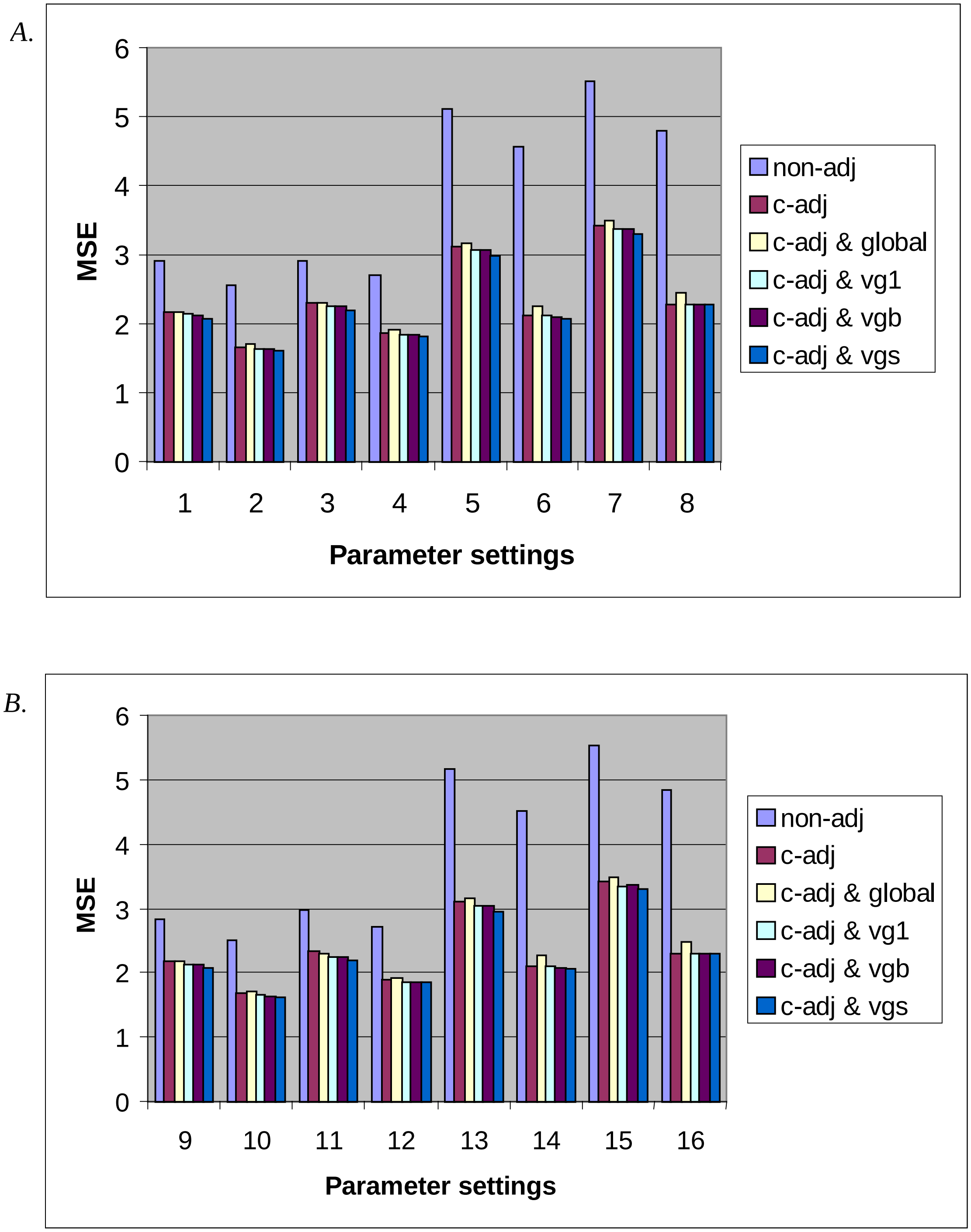


Figure 2.2.6 Mean square error (MSE) of full-sib genetic values (GV) from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 (A) test series and 8 (B) test series

Note: notation is the same as figure 2.2.3

Correlation between predicted genetic value and true value

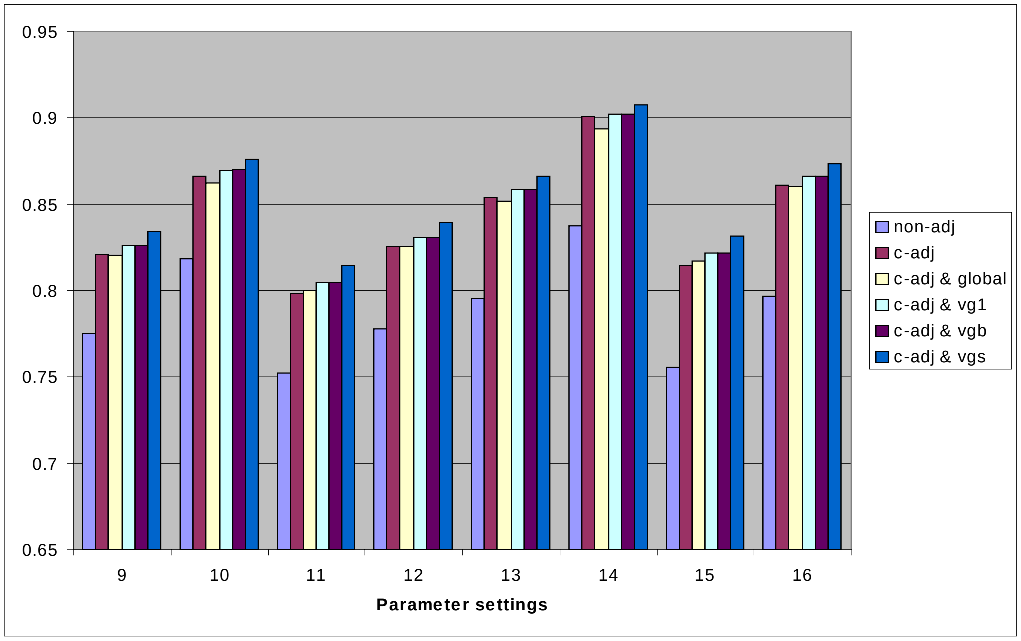
GCA

Correlation between BLUP and true value is probably the most important criterion. It predicts how much the expected genetic gain can be achieved in a tree breeding program.

Correlation of true GCA value and BLUP prediction ranged from .76 to .91 (Fig 2.2.7). Checklot mean adjustment for prediction improved correlation by more than .05. This significant improvement in correlation was consistent over all parameter structures. Additional gain in correlation was achieved by further adjusting GCA sample variance in c-adj&vg1 and c-adj&vgb. But use of global variance, i.e. average GCA variance over test region in adjustment (c-adj&global) does not necessarily increase correlation.

The same correlation pattern was observed for both simulated data with 4 test series and 8 test series. Genetic parameters affect the correlation in the following way: correlation of predicted full-sib GV ranged from 1.0 to 1.4 for simulated diallel data sets with heritability of .1 and from 1.1 to 1.8 for data sets with heritability of .2. Mean square error of GV prediction ranged from 1.6 to 3.0 for data sets of heritability .1 and from 2.1 to 5.5 for those of heritability .2. Similar to GCA prediction, both mean distance and mean square error did not differ much for both levels of test series number within a test region.

A.



B.

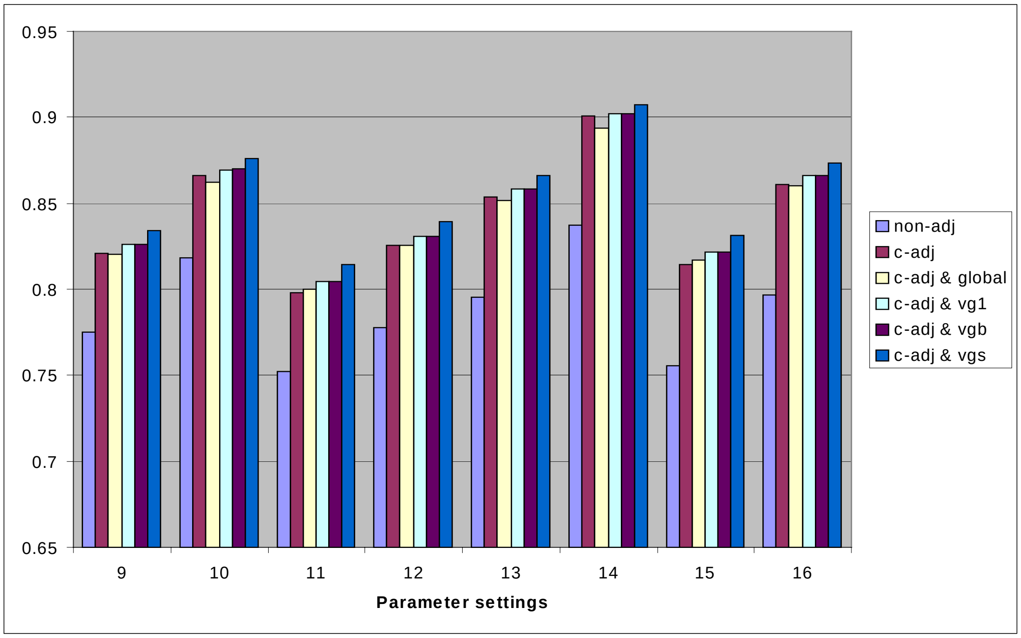


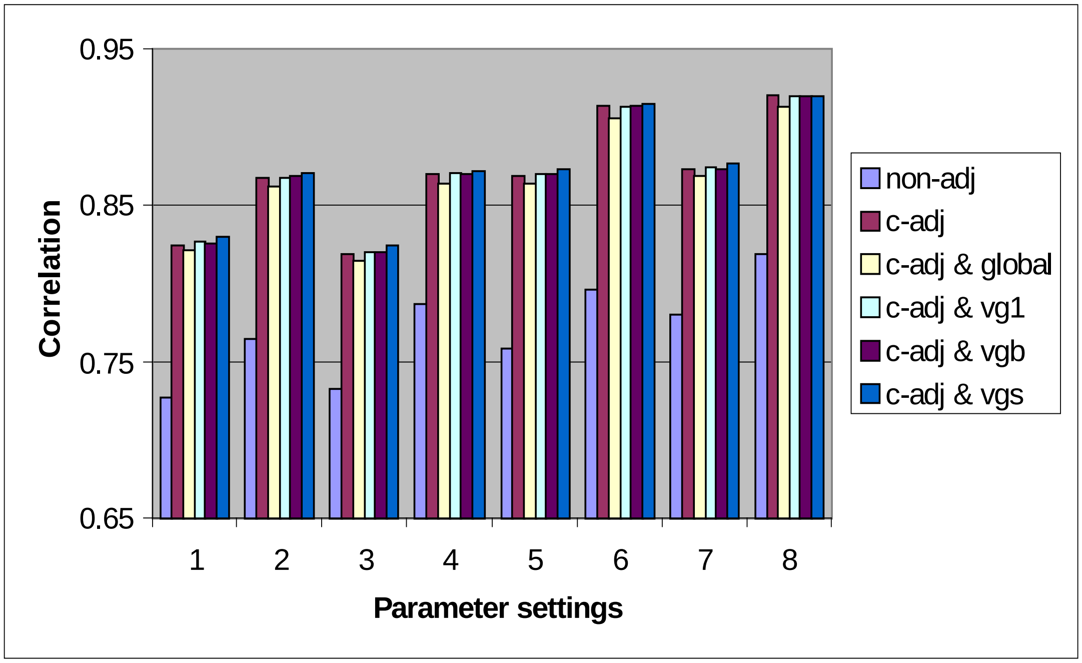
Figure 2.2.7 Mean correlation of parental GCA and prediction from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 (A) test series and 8 (B) test series

Note: notation is the same as figure 2.2.3

Full-sib family GV

Correlation between BLUP and true value for full-sib family GV has direct implications for tree breeding programs. By selecting on full-sib genetic value, both additive and non-additive genetic components in family mean variance can be captured (Mullin & Park, 1992). Genetic gain can be realized through mass controlled pollination of selected top full-sib families. It can also be combined with within family clonal selection to achieve the maximum genetic gain.

The range of correlation of true GV value and its BLUP prediction was from .73 to .93 (Fig 2.2.8). Regardless of parameter structure, checklot mean adjustment for prediction consistently improved correlation by more than .1. Similar to correlation of GCA, adjustment for GCA sample variance as in c-adj&vg1 and c-adj&vgb provided additional improvement in correlation, which stayed very close to theoretical limit. But the use of global variance in adjustment (c-adj&global) gave correlations lower than those from checklot adjustment alone.



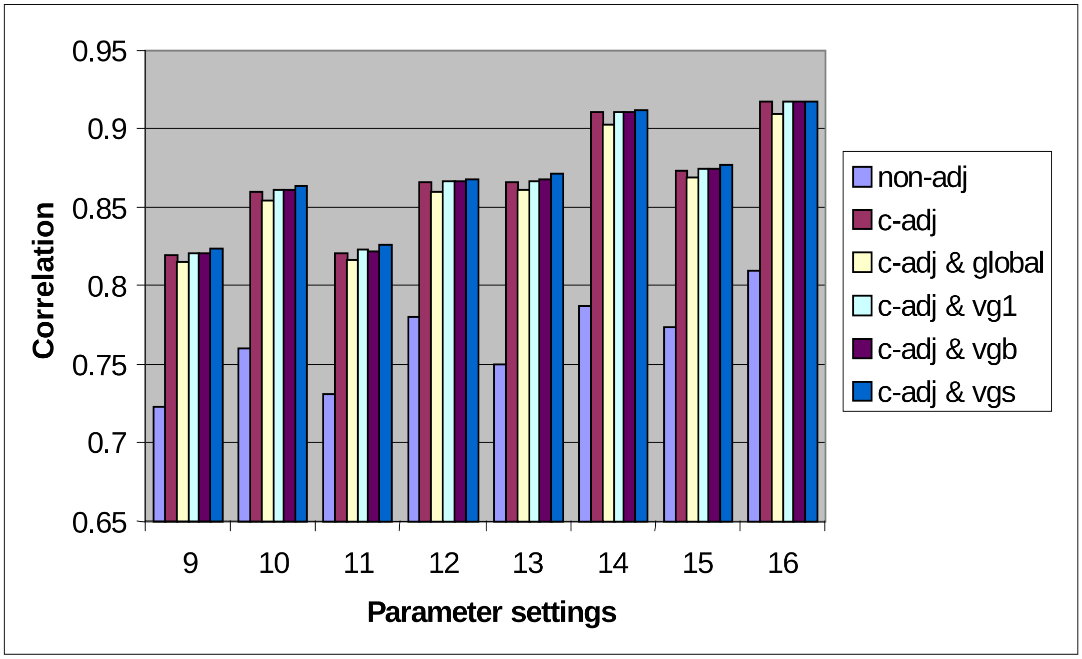


Figure 2.2.8 Mean correlation between full-sib true genetic value and prediction from BLUP analysis of simulated data sets (1000 for each parameter structure) across 4 (A) test series and 8 (B) test series

Note: notation is the same as figure 2.2.3

Conclusions

The importance of checklot adjustment in disconnected mating designs is often emphasized (Huber, 1990; van Buijtenen, J. P. & Bridgwater, 1986). The reason for the adjustment is usually attributed to heterogeneity of family or parent samples. In this study, we assume that parents of all test series in a test region are from the same base population, hence they share the same genetic variance and have no difference in mean performance other than sampling error. However, even under this ideal condition, sampling error alone can cause BLUP of different test series to be not comparable.

The best sample variance prediction in the class of linear combination of local variance estimates is developed in this study. This improved local variance estimate can be utilized to adjust both parental GCA and full-sib GV prediction. Since no specific mating design is assumed in the derivation, the interpretation of results can be extended to any disconnected mating designs, including both half-sib and full-sib design, in multiple test locations.

Results from simulation in this study showed that, for multiple disconnected diallel test series, checklot adjustment is critical to improve the prediction of genetic values obtained using BLUP analysis. Hence it can significantly improve the chances of selecting best performers based on GCA or full-sib GV. The significant improvement of prediction by using checklot adjustment is surprising but has important implication for future progeny testing and data analysis. Checklot adjustment alone can increase correlation more than .05 in GCA prediction and almost .1 in full-sib GV in all parameter settings. Although additional adjustments (GCA sample variance prediction) can be used as additional adjustment, the added improvement is rather limited beyond checklot adjustment due to the fact that the accuracy of checklot-adjusted prediction alone is already close to the theoretical limit.

Reference

[1] Borralho, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. In: Eucalypt plantations: improving fibre yield and quality. Proceedings of CRCTHF-IUFRO Conference. Hobart, Australia. p141-145.

[2] Huber, D. A. et al, Ordinary least squares estimation of general and specific combining abilities from half-diallel mating designs. *Silvae Genetica* 41(4-5): 263-273. 1992.

[3] Huber, D. A. Optimal mating designs and optimal tehcniques for analysis of quantitative traits in forest genetics. Ph. D. dissertation. University of Florida. 1993

[4] Li B., S.E. McKeand, A.V. Hatcher, and R.J. Weir. 1997. Genetic gains of second generation selections from the NCSU-Industry Cooperative Tree Improvement Program. Pro. 24th South. For. Tree Impr. Conf., p. 234-238. Orlando, Florida, June 9-12, 1997.

[5] Li, Bailian, S.E. Mckeand, R.J. Weir. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the south-eastern US. p. 164-68. In: Tree improvement for sustainable tropical forestry. Proceedings of QFRI-IUFRO Conference. Caloundra, Queensland, Australia. 1996.

[6] Littell, R.C. et al. SAS system for mixed models. SAS Institute Inc. Cary, NC. 1996

[7] Mullin, T.J., and Park, Y.S. Estimating genetic gains from alternative breeding strategies for clonal forestry. Can. J. For. Res. 22: 14-23. 1992

[8] SAS Institute Inc. SAS/STAT Software: Changes and enhancements (through release 6.11). Cary, NC. 1996.

[9] Searle, S.R., Casella, G., and Mcculloch, C.E. Variance components. John Wiley & Sons Inc., New York, 501p. 1992.

[10] Swallow, W.H. and Monahan J.F. Monte Carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components. Technometrics 26(1): 47-57, 1984.

[11] van Buijtenen J. P. & Burdon R. D. Expected efficiencies of mating designs for advanced generation selection. *Can. J. For. Res.* 20:1648-1663. 1990.

[12] van Buijtenen, J. P. & Bridgwater, F. Mating and genetic test designs. In: Advanced generation breeding of forest trees. Southern Coop. Series Bull. 309. *Louisiana Agr. Exp. Stn.* Baton Rouge, LA. pp5-10.

[13] White T. L. & Hodge, G. R. Best linear prediction of breeding values in a forest tree improvement program. Theor. Appl. Genet. 76:719-727. 1988.

[14] White, T. L. & Hodge, G. R. Predicting breeding values with applications in forest tree improvement. Kluwer Academic Pub., Dordrecht, The Netherlands. 367pp. 1989.

[15] Yanchuk, A. D. General and specific combining ability from disconnected partial diallels of coastal douglas-fir. *Silvae Genetica* 45(1):37-45 (1996).

[16] Zobel, Bruce&John Talbert. Applied forest tree improvement. John Wiley and Sons. New York. NY. 1984.

Chapter 3 Time trend of genetic parameter estimates in growth traits of loblolly pine

Abstract

Annual measurement from a total of 275 parents, 690 full-sib families from 23 diallel tests of loblolly pine (*Pinus taeda* L.) in Northern, Coastal and Piedmont test regions, was used to reveal time trend of genetic parameters through age 8. Variance components were estimated from mixed model analysis on growth traits (height, DBH and volume) and were used to derive genetic variance components, heritabilities, and age-age genetic correlations.

In all three test regions, dominance variance was found to be less than additive variance. The range of dominance variance fell within 20%~40% of total genetic variance for all traits. Different trends of heritability estimates were found for three test regions but the general trend was that heritability increased over time. The magnitude of heritabilities for DBH and volume was found to be comparable with the corresponding heritabilities for height.

Age-age genetic correlations of early height with 8-year volume increased significantly in the first 3-4 years then level off after that. DBH and volume had higher age-age correlation than height. The trends of heritabilities and age-age correlations indicate that the optimum selection age could be as early as 3 for height and as early as 4 for DBH or volume. Early selection with volume at age 4-5 could maximize genetic gain.

Key words: Diallel mating design; BLUP; heritability; Type B genetic correlation; age-age correlation; *Pinus taeda* L.

Introduction

Reliable estimation of genetic parameters is essential for predicting future gains and choosing appropriate breeding strategies in tree breeding programs. Many studies have been done for estimation of genetic variation in growth traits of loblolly pine (Li et al 1997; Li et al 1996; Balocchi et al 1993; Foster 1986; Lambeth ea al. 1983; Franklin 1979). However, most published results were from either unimproved or first-generation progeny tests and based their conclusions on rather small sample sizes of families with limited reliability and precision.

The genetic variation and parameters of the improved population of loblolly pine is still unclear. How do parameters change after one generation of improvement in breeding program? Would the time trend be different from the unimproved populations on well-tested sites? How do test series and breeding regions differ in genetic parameters? On what traits and when should selection be made to maximize genetic gain? With high quality and uniform progeny tests, a large family sample size would be critical in obtaining reliable estimates of genetic parameters. In addition, knowing the time trend of genetic parameters would be very useful in developing optimal breeding strategy for future selection.

The North Carolina State University - Industry Cooperative Tree Improvement Program (NCSU-ICTIP) has completed 44 years of loblolly pine tree improvement in the Southeastern United States. Through the first 2 cycles of breeding, testing and selection substantial genetic gains have been achieved. The cooperative's tree improvement program for loblolly pine has now moved into its third generation (Li *et al.* 1996, Mckeand et al. 1997). To further implement the third generation and future breeding plans, it is necessary to thoroughly analyze and evaluate current data accumulated from the first and the second-generation progeny tests of NCSU-ICTIP. In this study, twenty-three test series from the Early Diallel Measurement Study (EDMS) from 3 test regions were used in analysis. Annual measurements from a total of 275 parents and 690 full-sib families were used to estimate a comprehensive range of genetic parameters: additive and dominance genetic variance, appropriate heritabilities for various selection methods and age-age genetic correlation for growth traits and reveal the time trend of each parameter through age 8. The objectives of this study are to determine the magnitude and time trend of genetic parameters of growth traits of loblolly pine and to examine if loblolly pine populations in test regions differ in both magnitude and time trend of genetic parameters.

Materials

Mating design and field design

- Mating design and field design

The disconnected-half-diallel mating design was used to generate progenies for the second-generation breeding program at NSCU-ICTIP (Table 3-1). Each of two disconnected half diallels consisted of 6 parents and a total of 30 full-sib crosses. In each test, full-sib crosses were replicated over 6 blocks and planted in 6-tree row plots in each replication (Li *et al.* 1996). Four checklot families were also included in each test with two row plots in each replication.

Table 3-1. Mating design: disconnected diallel

	Diallel 1						Diallel 2					
F \ M	1	2	3	4	5	6	7	8	9	10	11	12
1		×	×	×	×	×						
2			×	×	×	×						
3				×	×	×						
4					×	×						
5						×						
6												
7								×	×	×	×	×
8									×	×	×	×
9										×	×	×
10											×	×
11												×
12												

- Regional tests

A total of 65 loblolly pine progeny tests in 15 test series with annual growth measurements were used for analysis. These tests can be grouped into the following three areas: 1) Virginia and northern North Carolina; 2) Atlantic Coastal Plains, Georgia and Lower Gulf; 3) Piedmont of Georgia, South Carolina and North Carolina and Upper Gulf.

These three test areas are referred to as Northern region, Coastal region and Piedmont region respectively. The above grouping of geographic areas is adopted for the 3rd cycle breeding plan by the loblolly pine breeding program at NCSU-ICTIP (NCSU-ICTIP, 1999). Each test region includes 4 to11 test series and each test series included 4 tests, which were planted in two years on two different sites. For every test region, there was at least one test series established in 1987 or earlier (McKeand 1987).

Data collection and growth curve correction

All progeny tests were measured annually up to age 8 for height and survival, and measurement for DBH started at age 4. Tree height was measured to the nearest centimeter, while DBH was measured to the nearest millimeter. Volume was calculated according to the following formula:

$$\mathrm{Vol} = 0.03371 + (0.00196128 \cdot \mathrm{DBH}^2 \cdot \mathrm{HT})$$

Richards' function was used to edit the data to correct irregularities in the data, such as leaders broken by adverse weather, damaged by insects or diseases and measurement or data recording errors. This flexible growth model permits each individaul tree to have its own growth curve (Richards, 1959; Balocchi, 1990). The form of this equation is shown as follows:

$$Y = A(1 - e^{-bX})^c$$

Where

Y = measurement of a certain character (e.g. height)

A = the upper limiting value for the character e = the base of natural logarithums

b & c = shape parameters which determine the shape of the curve along the time axis

X = age in years

All further analysis will be performed on the predicted values of this nonlinear model.

Linear Statistical model

For each test series, the following linear model is used to estimate variance components.

$$Y_{ijoklm} = \mu + T_i + B_{j(i)} + D_o + G_{k(o)} + G_{l(o)} + S_{kl(o)} + TG_{ik(o)} + TG_{il(o)} + TS_{ikl(o)} + P_{ijokl} + E_{ijoklm}$$

Where Y_{ijoklm} is the *l*th observation of the *i*th block for the *jk*th cross

μ is the overall mean

T_i is the *i*th fixed test environment (location) effect

$B_{j(i)}$ is the fixed effect of *j*th block within *i*th test

D_o is the *o*th fixed diallel effect

$G_{k(o)}$ or $G_{l(o)}$ is the random general combining ability effect of the *k*th female or *l*th male within *o*th diallel(*k*≠*l*) ~NID N(0, σ^2_g);

$S_{kl(o)}$ is the random specific combining ability effect of *k*th and *l*th parents within *o*th diallel(*k*≠*l*) ~NID N(0, σ^2_s)

$TG_{ik(o)}$ or $TG_{il(o)}$ is the random test by female or male GCA interaction ~NID N(0, σ^2_{gt})

$TS_{ikl(o)}$ is the random test by SCA interaction ~NID N(0, σ^2_{st})

P_{ijokl} is the random plot effect for the *kl*th cross within *o*th diallel in the *i*th block ~NID N(0, σ^2_p)

E_{ijoklm} is the random within plot error term ~NID N(0, σ^2_e)

All random effects are assumed to be independent of each other. The model is a typical mixed model and is analyzed in SAS MIXED procedure by creating dummy variables for GCA effects (see chapter 1).

Genetic variance and heritability estimation

The genetic variance components are calculated with SAS PROC MIXED procedure (Chapter 1). Genetic and other parameters are defined and calculated using the variance estimates as follows.

Additive and nonadditive genetic variance: $\sigma^2_A = 4\sigma^2_g$ $\sigma^2_D = 4\sigma^2_s$

Narrow sense individual heritability: $h^2=4\sigma^2_g /(2\sigma^2_g+ \sigma^2_s+ 2\sigma^2_{gt}+ \sigma^2_{st}+ \sigma^2_p +\sigma^2_e)$

Broad sense individual heritability: $H^2=4(\sigma^2_g +\sigma^2_s)/(2\sigma^2_g+ \sigma^2_s+ 2\sigma^2_{gt}+ \sigma^2_{st}+ \sigma^2_p +\sigma^2_e)$

Half-sib mean heritability:

$$h^2{}_{HS} = \frac{\sigma^2_g}{\sigma^2_d + \frac{p\sigma^2_g}{p-1} + \frac{\sigma^2_s}{p-1} + \frac{p\sigma^2_{gt}}{t(p-1)} + \frac{\sigma^2_{st}}{t(p-1)} + \frac{\sigma^2_p}{bt(p-1)} + \frac{\sigma^2_e}{bnt(p-1)}}$$

Full-sib family mean heritability (narrow sense):

$$h^2{}_{FS} = \frac{2\sigma^2_g}{\sigma^2_d + 2\sigma^2_g + \sigma^2_s + \frac{2\sigma^2_{gt}}{t} + \frac{\sigma^2_{st}}{t} + \frac{\sigma^2_p}{bt} + \frac{\sigma^2_e}{bnt}}$$

Full-sib family mean heritability (broad sense):

$$H^2{}_{FS} = \frac{2\sigma^2_g + \sigma^2_s}{\sigma^2_d + 2\sigma^2_g + \sigma^2_s + \frac{2\sigma^2_{gt}}{t} + \frac{\sigma^2_{st}}{t} + \frac{\sigma^2_p}{bt} + \frac{\sigma^2_e}{bnt}}$$

Narrow sense within full-sib family heritability:

$$h^2{}_{WFS} = \frac{2\sigma_g^2}{\frac{2(t-1)\sigma_{gt}^2}{t} + \frac{(t-1)\sigma_{st}^2}{t} + \frac{(bt-1)\sigma_p^2}{bt} + \frac{(bnt-1)\sigma_e^2}{bnt}}$$

Broad sense within full-sib family heritability:

$$H^2{}_{WFS} = \frac{2\sigma_g^2 + 3\sigma_s^2}{\frac{2(t-1)\sigma_{gt}^2}{t} + \frac{(t-1)\sigma_{st}^2}{t} + \frac{(bt-1)\sigma_p^2}{bt} + \frac{(bnt-1)\sigma_e^2}{bnt}}$$

The estimates of variance components and genetic parameters were averaged across all test series of in each test region to compare among traits and regions.

Genetic correlation and type B genetic correlation estimation

Using the same variance estimation procedure (as in 1.), the genetic correlation are estimated on the sum of the values of two variable and then using the following formula:

$$r_{Gxy} = \frac{\sigma_{gxgy}}{\sqrt{\sigma_{gx}^2 \cdot \sigma_{gy}^2}} = \frac{\left[\sigma_{gx+gy}^2 - \sigma_{gx}^2 - \sigma_{gy}^2\right]/2}{\sqrt{\sigma_{gx}^2 \cdot \sigma_{gy}^2}}$$

Where

σ_{gxgy} is the genetic covariance between two traits,

σ_{gx+gy}^2 is the variance estimate of the sum of two additive genetic values of two traits.

Since the data is generally well balanced in terms of crosses and missing plots and the estimation of σ_e^2 is not of particular interest, plot means are instead used to in analysis to obtain $\sigma^2{}_{gx+gy}$.

The type B genetic correlation is calculated following the formula given by Yamada (1962), which is the well-accepted measurement of genetic by environmental effects (Burdon, 1977).

$$r_B = \sigma^2_A/(\sigma^2_A+\sigma^2_{AE}) = \sigma^2_g/(\sigma^2_g+\sigma^2_{gt})$$

Results

Time trend of genetic parameters

Genetic variance estimates

The genetic variance increased over time and had similar trend for three test regions (Fig. 3-1). Additive genetic variance (σ^2_A) was small but increased in a linear fashion for HT and DBH over time. Exponential increase was observed for volume additive genetic variance. Dominance genetic variance (σ^2_D) also increased over time and had generally the same pattern as the additive genetic variance, but they were less predictable than additive component in Coastal and Piedmont regions.

Three test regions differed in magnitude of genetic variance components. Coastal region had the highest variance component estimates. For height at age 8, additive genetic variance was higher (1.6) in Coastal region than those in the Northern and Piedmont region (.5 and .7 respectively). The same order in magnitude of genetic variance components was observed for both DBH and volume.

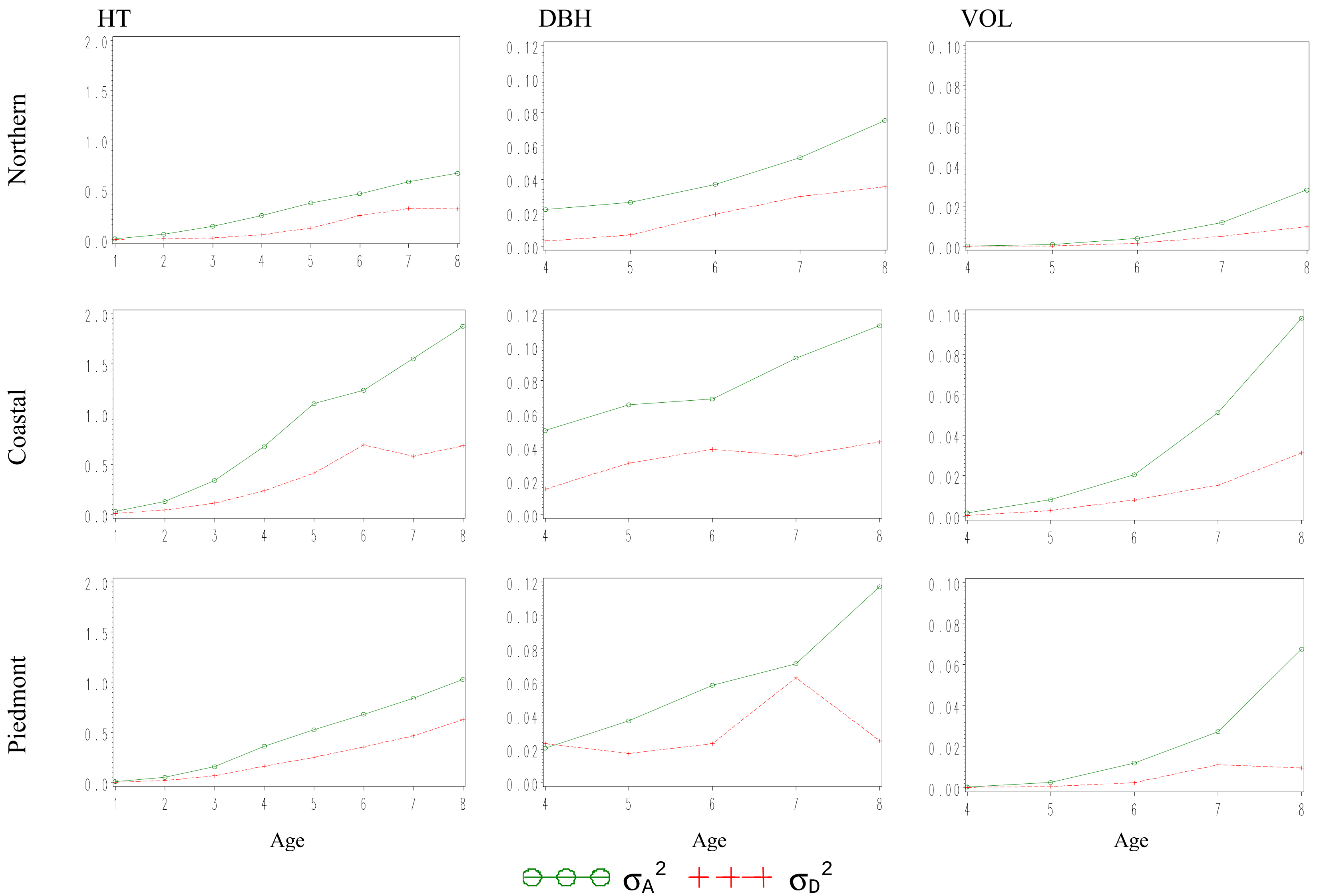


Figure 3-1. Time trend of additive (σ^2_A) and dominance (σ^2_D) variance component estimates for trait height (HT), diameter ad breast height (DBH) and Volume (VOL) during early growth period in 3 test regions (Northern, Coastal and Piedmont)

Heritability

Narrow sense individual heritability estimates ranged within .05~.20. Non-additive variance added about 30% into the broad sense heritability, resulting in a range of .10~.30. The family heritabilties ranged from .5 to .8 (Fig. 3-4).

While the general trend of heritability estimates was increasing over time and are similar for three growth traits, the detailed shape of the curves varied for different heritabilities and test region. In the Northern region all heritability estimates did not vary significantly with increasing age for all three traits. Slight increases were observed over time for individual heritabilities, while family heritabilities were rather stable over ages (Fig. 3-4, 3-5).

In Coastal and Piedmont test region, individual heritabilities for height increased over time until age 4 or 5, then plateaued in Piedmont region declined in Coastal region (Fig. 3-4). The corresponding family heritability of the same trait follows similar but less marked changes as the individual heritabilities (Fig. 3-5).

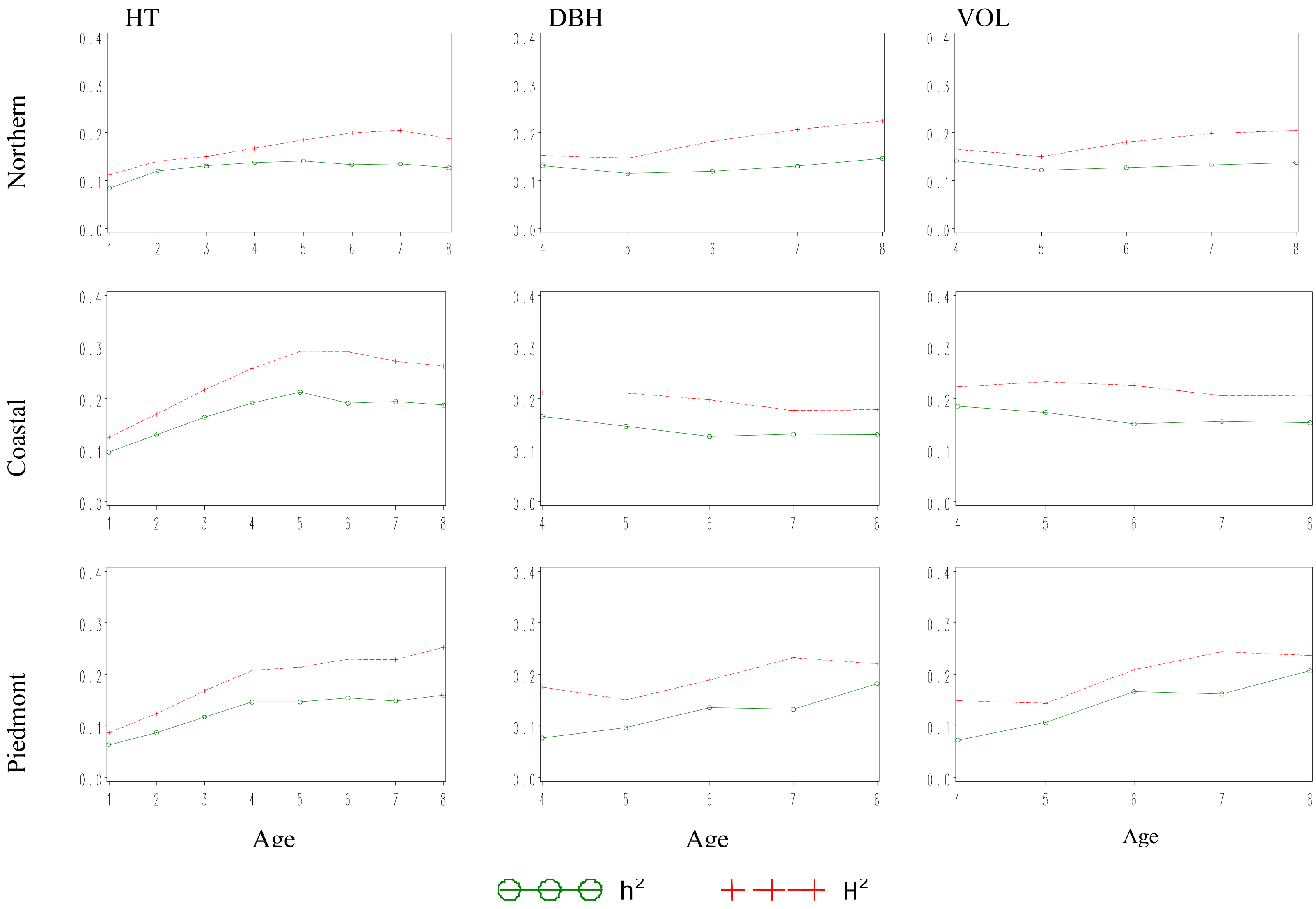


Figure 3-4. Time trend of narrow sense (h^2) and broad sense (H^2) individual heritabilities for trait height (HT), diameter at breast height (DBH) and Volume (VOL) during early growth period in 3 test regions (Northern, Coastal and Piedmont)

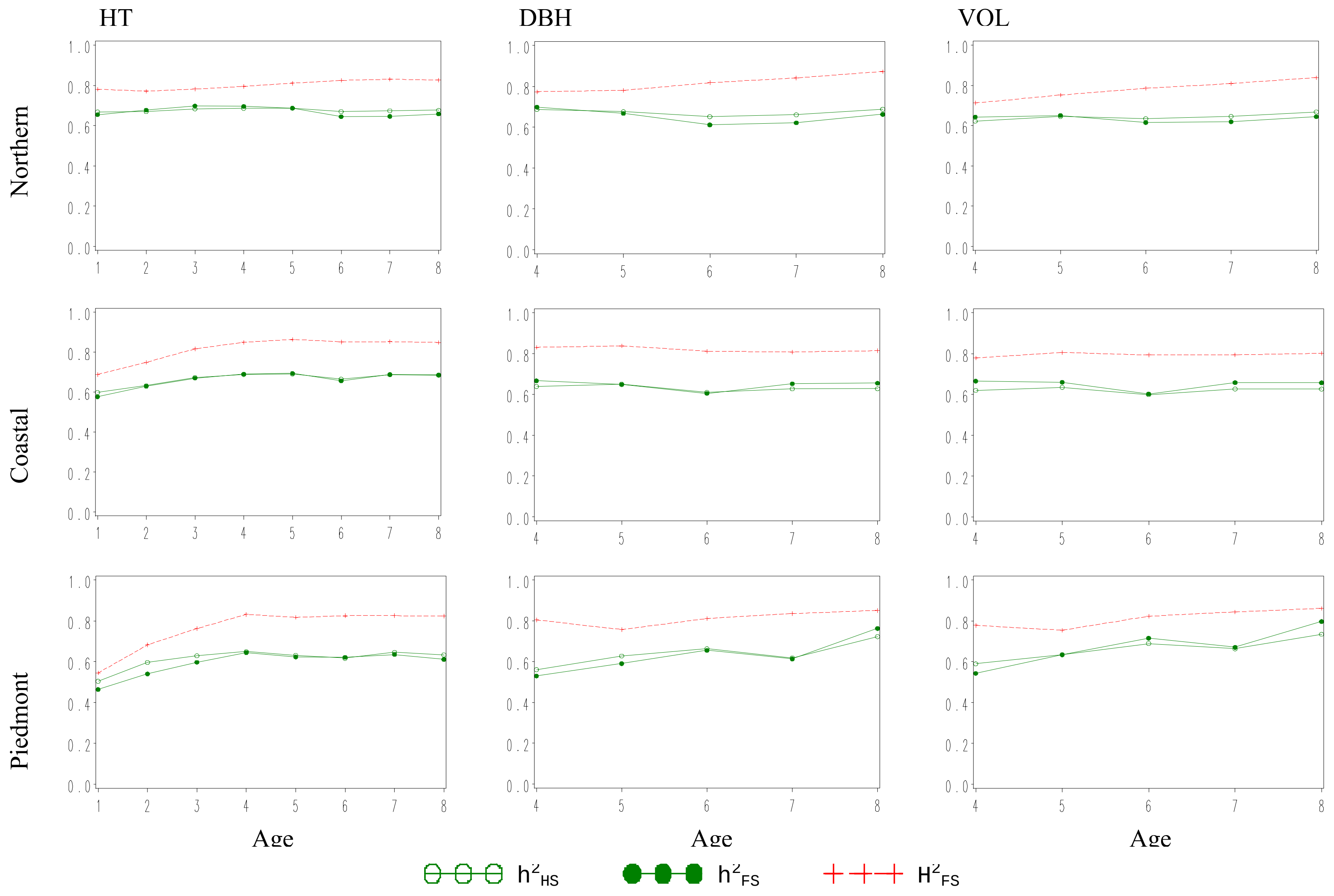


Figure 3-5. Time trend of family mean heritabilities for trait height (HT), diameter at breast height (DBH) and Volume (VOL) during early growth period in 3 test regions (Northern, Coastal and Piedmont)

Type B genetic correlation

The type B genetic correlation for height ranged from .65 to .93, with most values between .7 and .85. Similarly, DBH had high correlation across test region and ages, ranging from .68 to .83. Volume had slightly lower correlations, with its range from .62 to .78 (Table 3-2).

In the Northern region, the type B genetic correlation is consistently high (larger than .7) across all ages for height and DBH, showing weak genetic by environment interaction effect within the test series for all ages. Correlation for volume starts was relatively low (.62) then increased gradually to .71 at age 8. In other two test regions, type B correlation for height was high for most ages except at age one. Correlation for DBH was slightly lower than .7 after age 5 in Costal region. Similar to Northern region, Piedmont region had constant high correlation (larger than .7) across ages for DBH. Unlike Northern region, volume correlation for Costal region started as high as .73 then goes down slightly to .67 at age 8, while in Piedmont region volume correlation maintained a high level (larger than .7) across ages except for .64 at age 5.

Table 3-2. Averaged type B genetic correlation for growth traits in three test regions

	Northern region			Coastal region			Piedmont region		
AGE	HT	DBH	VOL	HT	DBH	VOL	HT	DBH	VOL
1	0.7444	-	-	0.6849	-	-	0.6540	-	-
2	0.7124	-	-	0.6901	-	-	0.7348	-	-
3	0.7412	-	-	0.7873	-	-	0.8444	-	-
4	0.7552	0.7546	0.6161	0.8117	0.7980	0.7296	0.9287	0.7243	0.7750
5	0.7688	0.7563	0.6399	0.8192	0.8257	0.7412	0.8361	0.7443	0.6358
6	0.7892	0.7649	0.6685	0.7692	0.7133	0.6800	0.8304	0.7750	0.7274
7	0.8102	0.7846	0.7092	0.8012	0.6843	0.6676	0.8217	0.7751	0.7370
8	0.7926	0.8302	0.7809	0.7899	0.6816	0.6660	0.7853	0.7951	0.7922

The type B correlation increased slightly over time for height. But for DBH and volume, the general trend over time was not clear, partly because of short time period from age 4 to 8. Overall type B genetic correlation was fairly stable over time and test region for all three traits within the measurement time frame.

Genetic correlation

Age-age genetic correlation for the same trait

The genetic correlations between age 8 and previous ages of the same trait are shown in Fig 3-10. As expected, all correlation increased over time and approached unity at age 8. The range for correlation of height was from .3 at age 1 in Northern region and went up to .9 at age 5. The lowest correlation for DBH and volume was around .6 at age 4 in Coastal region. High correlations were observed at very early growth stage for all 3 regions.

Early height had the highest genetic correlation among three traits in Northern region and Piedmont region, followed by volume and DBH. On the contrary, in Coastal region the genetic correlation for volume slightly exceeded that for height. Correlation for DBH was slightly lower than other two traits in Coastal region.

In Northern and Coastal regions, the genetic correlation for early height rose rapidly from age 1 to age 3. At age 3, the correlation exceeded or was close to 0.8 and continued to increases gradually. The age-age genetic correlation for DBH and volume starts were relatively low at early ages but then increased similar to height in Northern region, while correlation was initially high (.9) at age 4 and stayed high as height in Coastal region. In Piedmont region, the correlation was high (.9) as age 2 and stayed high through age 8. The correlation for DBH was slightly lower than that for volume and height in Piedmont region than other two regions.

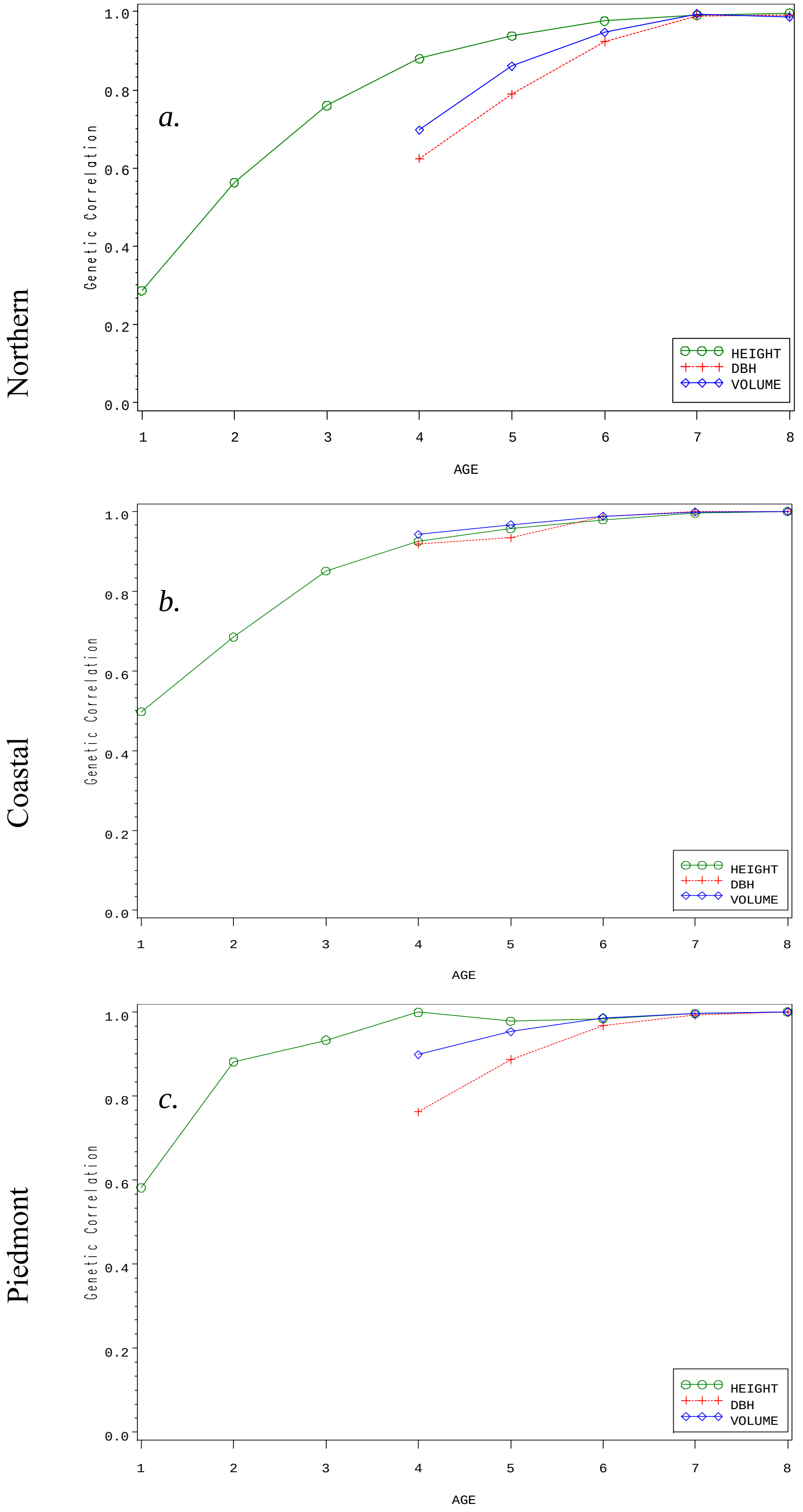


Figure 3-10 Comparison of age-age correlation of growth traits for 3 test regions: Northern (a), Coastal (b) and Piedmont (c)

Age-age genetic correlation with volume at age 8

The genetic correlations of early height, DBH and volume with age-8 volume were presented in Fig 3-11. Correlations generally increased with time for all traits. The correlation for trait height increased from .23 to .74 in Northern region, from .55 to .85 in Coastal region and from .74 to 88 in Piedmont region. For trait DBH, correlations with volume were higher than those for height in all three regions. Juvenile volume had the highest correlations with 8-yr volume as expected in all three regions.

In Coastal region and Piedmont region, the difference of average genetic correlation between height and DBH or volume became smaller than that in Northern region. Particularly in Piedmont region, correlation at age 4 for height was even higher than DBH. However, it must be noted here that genetic correlation of tree height with VOL8 had more variation among test series. It exceeded 1 in some test series of Piedmont region, which resulted in a slightly higher correlation than other two regions. The shape of trend over time for trait height was also different for any of other two traits, in that it had higher correlation in the middle of measurement time duration. It peaked at age 5 for both Northern region and Coastal region, and had two mode, age 3 and 5, in Piedmont.

Volume had the highest correlation among three traits in Northern region and Coastal region, followed by DBH and height. Volume still had its best correlation in Piedmont region, but height seemed to have the higher genetic correlation in earlier years. Visibly, it exceeded DBH at age 4 when measurements were available for comparison among three traits. Finally, three test regions also differed in the shape of curves shown in the Figure 3-11. The increasing rate of genetic correlation in early ages was larger in Northern region than in other two regions. The later two regions differed in that height had more change over time in Coastal region while DBH and volume had more age-age difference in Piedmont region.

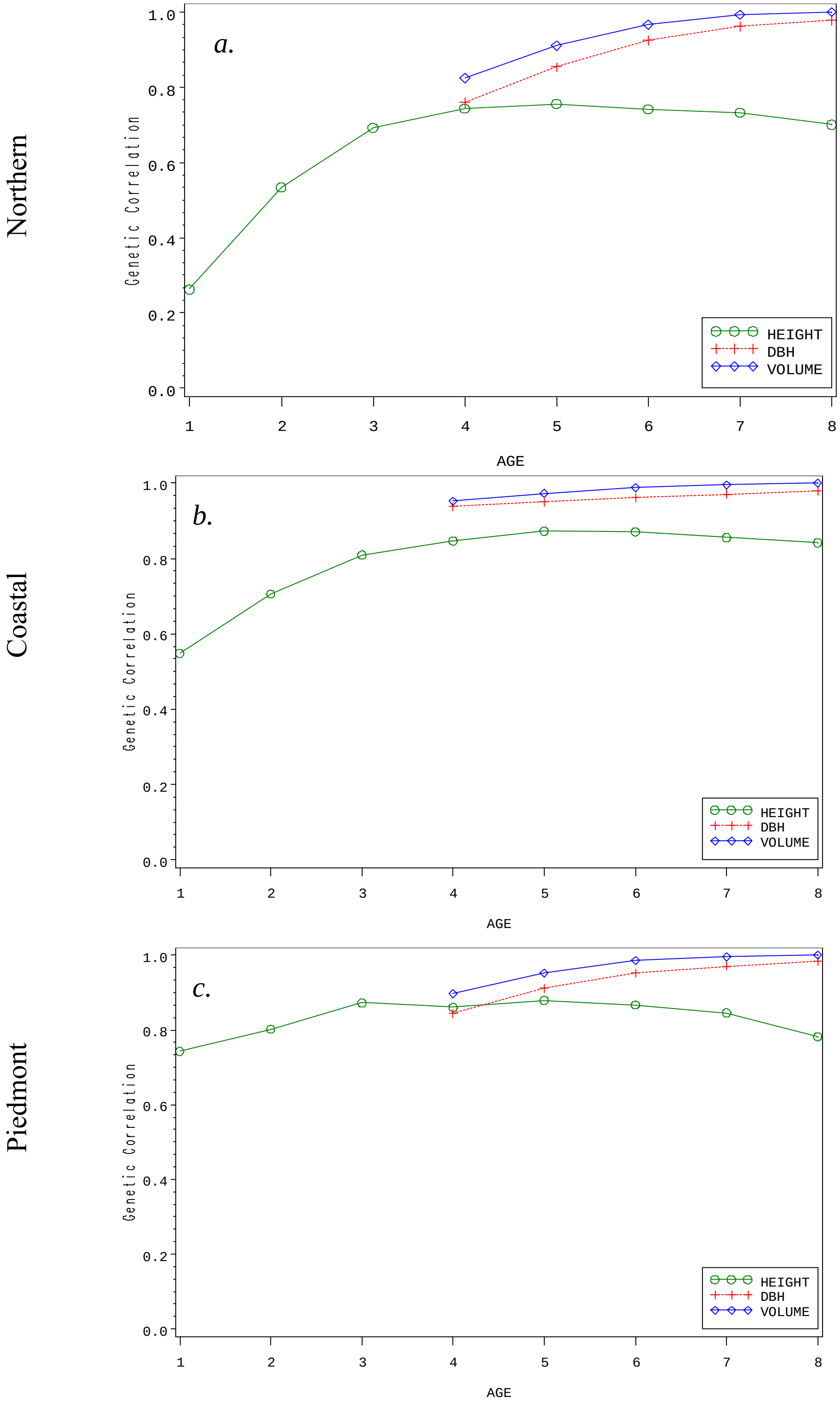


Figure 3-11. Comparison of correlation with age 8 volume of growth traits for 3 test regions: Northern (a), Coastal (b) and Piedmont (c)

Discussion

The genetic variance components increased over time and the trends are similar for all test regions. The increase for both height and DBH was almost linear with age, while the increase for volume was exponential. Three test regions differed in magnitude of variance components. Coastal population had the highest variance components, followed by Piedmont source and Northern source. This may be due to difference in growth rate of different regions. Loblolly pine trees have fast growth rate in Coastal region. Northern population is known to be cold-hardy but with relatively slow growth rate. Piedmont source is also cold-hardy with better stem form and its growth rate is in the middle of two other regions (NCSU-ICTIP, 1999).

Unlike the results from Balocchi (1992), which showed that dominance variance exceeded additive variance at ages before 12, we found that the percentage of dominance variance in the total genetic variance only varied from 10% to 40% for all three growth traits across test regions. On average, the σ^2_D was around 30% of the total genetic variance for all three growth-traits across all regions. In addition, the percentage of dominance was the highest at around age 6 or 7 for Northern and Coastal populations. In Piedmont region, however, no clear pattern was observed. This time trend of genetic variance generally agrees with the results from ealier study on the similar diallel tests (Li, 1996).

Typical heritability estimates for growth traits were observed in this study. Unlike earlier results for Northern population (Annual report, 1999), heritability estimates showed different time trends for Coastal region and Piedmont region. For tree height, heritabilities (includes individual, family heritability and broad sense heritabilities) were all found to increase over time. Although large variation of heritability existed among test series, when averaged over entire region we were able to find a general increase pattern for tree height in both regions: heritabilities increased from age 1 to age 4 and then stabilized after that. For DBH and volume, time trend resembled but was more apparent than that of tree height in the period from age 4 to age 8. Most importantly, unlike other studies (Svensson, 1999; Lambeth, 1980), the magnitude of heritabilities for DBH and volume were found to be comparable with the corresponding heritabilites for height.

The type B genetic correlation was generally high across all test regions. This confirms early results, which showed that loblolly pine has little genotype by environment interaction and have high family stability in performance across a wide geographic area (Li and McKeand 1989, McKeand et al. 1990).

Genetic correlation is highly derivative estimate and is subject to large sample error drift. In our study, the sample size of parents is fairly large, ranging from 48 (region 1) to 120 (Coastal region). Hence we would expect that the result should be reliable and produce useful information for selection and breeding in the future.

General high age-age correlation for height indicates if our goal is to improve 8-year height, early selection can be considered at age 4, or even at age 3 in case of Coastal region. This also confirms the results from Li (1996) that early height is a good prediction of 8-yr height growth.

If our selection is for 8-yr volume, correlation for DBH and volume were better than height in all regions except early ages for Piedmont population. These results indicated that although genotypic height performance at age 3-4 is a good indicator of performance at age 8, selecting on height alone might not be most efficient for volume. DBH and volume at age 4-5 could be considered to achieve maximum genetic gain.

Conclusions

In all three test regions, additive genetic variance was found to be more important than dominance variance. The relative proportion changed slightly over time but fell within 20%~40% of total genetic variance for all traits. This result, together with earlier analysis on the similar diallel tests (Li, 1996), indicate that selected loblolly pine populations may have different genetic structure and time trends from unselected populations (Balocchi 1992).

Age-age genetic correlations of early height with 8-year volume increased significantly in the first 3-4 years and then leveled off. DBH and volume had higher age-age correlation than height. The trends of heritabilities and age-age correlations indicate that the optimum selection age could be as early as 3 for height and as early as 4 for DBH or volume. Early selection with volume at age 4-5 could maximize genetic gain.

Reference

[1] Balocchi, C.E. et al. Age trends in genetic parameters for tree height in a nonselected popultaion of loblolly pine. *For. Sc.* 39:231-251. 1993.

[2] Borralho, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. In: Eucalypt plantations: improving fibre yield and quality. Proceedings of CRCTHF-IUFRO Conference. Hobart, Australia. p141-145.

[3] Burdon, R.D. Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. Silvae Genet. 26:168-175. 1977

[4] Foster, G.S. Trends in genetic parameters with stand development and their influence on early selection for volume growth in loblolly pine. Forest Sci. 32: 4. 1986.

[5] Huber, D.A. Optimal mating designs and optimal tehcniques for analysis of quantitative traits in forest genetics. Ph. D. dissertation. University of Florida. 1993

[6] Li, B., McKeand, S.E, Hatcher, A.V., and Weir, R.J. 1997. Genetic gains of second generation selections from the NCSU-Industry Cooperative Tree Improvement Program. Pro. 24th South. For. Tree Impr. Conf., p. 234-238. Orlando, Florida, June 9-12, 1997.

[7] Li, B., Mckeand, S.E, and Weir, R.J. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the south-eastern US. p. 164-68. In: Tree improvement for sustainable tropical forestry. Proceedings of QFRI-IUFRO Conference. Caloundra, Queensland, Australia. 1996.

[8] Littell, R.C. et al. SAS system for mixed models. SAS Institute Inc. Cary, NC. 1996

[9] McKeand, S.E. Optimum age for family selection for growth in genetic tests of loblolly pine tree. For. Sci. 43:400-411.1988.

[10] Richards, F. Aflexible growth function for empirical use. Jour. of Exp. Botany. 1959.

[11] SAS Institute Inc. SAS/STAT Software: Changes and enhancements (through release 6.11). Cary, NC. 1996.

[12] Schaffer, H.G. and Usanis, R.A. General least square analysis of diallel experiments. A computer program DIALL. N.C. State Univ., Genet. Dept., Res., Rep. 1, 1969.

[13] Searle, S.R., Casella, G., and Mcculloch, C.E. Variance components. John Wiley & Sons Inc., New York, 501p. 1992.

[14] Swallow, W.H. and Monahan, J.F. Monte Carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components. Technometrics 26(1): 47-57, 1984.

[15] van Buijtenen, J.P. and Burdon, R. D. Expected efficiencies of mating designs for advanced generation selection. Can. J. For. Res. 20:1648-1663. 1990.

[16] van Buijtenen, J.P. and Bridgwater, F. Mating and genetic test designs. In: Advanced generation breeding of forest trees. Southern Coop. Series Bull. 309. Louisiana Agr. Exp. Stn. Baton Rouge, LA. pp5-10.

[17] White, T.L. and Hodge, G. R. Best linear prediction of breeding values in a forest tree improvement program. Theor. Appl. Genet. 76:719-727. 1988.

[18] White, T.L. and Hodge, G. R. Predicting breeding values with applications in forest tree improvement. Kluwer Academic Pub., Dordrecht, The Netherlands. 367pp. 1989.

[19] Yamada, Y. Genoype by environmental interaction and genetic correlation of the same trait under different environments. Jap. J. Genet. 37:498-509.

[20] Yanchuk, A.D. General and specific combining ability from disconnected partial diallels of coastal douglas-fir. Silvae Genetica 45(1):37-45 (1996).

[21] Zobel, B.J. and Talbert, J.T. Applied forest tree improvement. John Wiley and Sons. New York. NY. 1984.

Chapter 4 Optimal Selection Strategy of Loblloly Pine

Abstract

Time trend of genetic parameters and selection efficiencies of height and volume were examined for three test regions of Loblloly pine. Reliable genetic parameters and breeding values for height and volume were estimated or predicted from analysis of annual diallel progeny tests. Genetic gain in year-8 volume was predicted for various selection methods at age 6 for both traits. Selection on volume was found to yield more gain than selection on height. Among test regions, Coastal population had the greatest correlated response, followed by Piedmont population and Northern population. Family plus within family selection based on total genetic component can capture the most genetic gain. For all selection methods, additional gain (10-40%) can be achieved by capturing non-additive genetic component.

The criterion of selection efficiency was formulated as the ratio of either gain per year (SE1) or present value (SE2) between indirect selection and direct selection. Optimal selection ages were determined for various selection methods. Compared with direct selection on volume of age 8, earlier selection appeared to be more efficient in most of selection methods. Family selection can be performed as early as age 2 or 3 for trait height and at age 4 for DBH or volume, which is the earliest measurement year in this study. Combined selection (family plus within family) was highly efficient at age 3 or age 4. The results from BLUP selection indicated similar result except that genetic gains estimated from BLUP selection were higher than those from family selections based on heritability. Additional gain can be achieved by selecting on full-sib genetic value.

Key words: Diallel mating design; best linear unbiased prediction (BLUP); general combining ability (GCA); breeding value; selection efficiency; selection age; *Pinus taeda* L.

Introduction

In any tree breeding program, the accurate assessment of genetic parameters is critical for predicting future gains and developing appropriate breeding strategies. As the most important commercial tree crop in Southern US (Mckeand et al. 1997; Lantz 1987), genetic parameter estimates of Loblolly Pine (*Pinus taeda* L.) were reported from the various studies (McKeand 1988; Foster 1986; Lambeth et al. 1983; Franklin 1979). These studies indicate that early selection on growth traits could be effective. But these studies based their conclusion on rather rough trends of genetic parameters over time. In addition, in most studies small family sample size limited the reliability and precision of estimates.

Balocchi et al. (1993) studied selection efficiency from the analysis of data from an unimproved population of Loblolly Pine. Time trends in genetic parameters for height indicated that, if a single measurement is used, measurement at age 6 and selection one year later would maximize the gains per year. However, genetic parameters and trends over time may be different with improved loblolly pine population that were generated with different mating designs and field layouts. The NCSU-ICTIP has been using half-diallel established on very uniform land with early diallel measurement study (EDMS). In this study, the well-balanced data from second-generation genetic testing with a disconnected diallel mating design is used to estimate genetic parameters at early ages more accurately and precisely. With better estimates of parameters, time trends of selection efficiency are reevaluated for different selection methods in order to get maximum genetic gains per time unit.

Besides selection efficiency study based on population genetic parameter estimates, genetic prediction (GCA or breeding value) from BLUP analysis is also used in this study for both half-sib and full-sib family selection. BLUP has been widely used because of its advantages over other gain prediction or selection methodologies (White and Hodge, 1988, 1989; Borralho, 1995). Hence selection strategy based on BLUP would be more useful from the practical point of view. In this study, family selection on BLUP of GCA and full-sib genetic value and breeding value will be conducted over all applicable ages in order to compare gains and selection ages with prediction based on population parameters.

In many studies, non-additive variance is an important part of genetic variance at the early growth stage of loblolly pine (Balocchi, 1994; McKeand et al. 1986; Foster et al., 1986). Assessment of its impact on genetic gain improvement will be valuable in making decisions to develop breeding strategies such as mass production of crosses and vegetative propagation. Non-additive variance component captured by using broad sense heritabilities or in case of BLUP simply adding SCA prediction will be assessed to achieve maximum genetic gain via vegetative propagation techniques.

Material and Methods

Mating design and field design

- Mating design and field design

A total of 275 parents, 690 full-sib families from 23 second-generation Loblolly pine diallel test series were included in this study. These tests were grouped into the following three regions (see chapter 3):

1. Northern region: Virginia and northern North Carolina;

2. Coastal region: Coastal Plains of South Carolina, Georgia and Lower Gulf;

3. Piedmont region: Piedmont of Georgia, South Carolina and North Carolina.

There were 4~11 test series in each test region. Each test series included 4 tests, planted over 2 years in 2 locations. In each test series, a disconnected-half-diallel mating design was used to generate progenies (Table 3-1). Each of two disconnected half diallels of 6 parents produced 15 full-sib crosses. In each test, full-sib crosses from two disconnected diallels were replicated over 6 blocks and planted in 6-tree row plots (Li *et al.* 1996). Four checklot families were also included in each test with 2 row plots in each replication.

All four tests in each test series were measured annually for height through eighth year. DBH was measured annually from age 4 through age 8.

Linear model and genetic parameter estimation

For each test series, the following linear model was used to estimate variance components.

$$Y_{ijoklm} = \mu + T_i + B_{j(i)} + D_o + G_{k(o)} + G_{l(o)} + S_{kl(o)} + TG_{ik(o)} + TG_{il(o)} + TS_{ikl(o)} + P_{ijokl} + E_{ijoklm} \quad (1)$$

Where Y_{ijoklm} is the lth observation of the ith block for the jkth cross

μ is the overall mean

T_i is the ith fixed test environment (location) effect

$B_{j(i)}$ is the fixed effect of jth block within ith test

D_o is the oth fixed diallel effect

$G_{k(o)}$ or $G_{l(o)}$ is the random general combining ability effect of the jth female or kth male within oth diallel(j≠k) ~NID N(0, σ^2_g);

$S_{kl(o)}$ is the random specific combining ability effect of ith and kth parents within oth diallel(j≠k) ~NID N(0, σ^2_s)

$TG_{ik(o)}$ or $TG_{il(o)}$ is the radom test by female or male GCA interaction ~NID N(0, σ^2_{gt})

$TS_{ikl(o)}$ is the radom test by SCA interaction ~NID N(0, σ^2_{st})

P_{ijokl} is the random plot effect for the klth cross within oth diallel in the ith block ~NID N(0, σ^2_p)

E_{ijoklm} is the radom within plot error term ~NID N(0, σ^2_e)

All random effects are assumed to be independent of each other. The model is a typical mixed model and was analyzed by using the new mixed analytical method described in the first chapter, i.e. utilizing SAS PROC MIXED by creating dummy variables for GCA effects. REML is chosen as the model fitting method, as it was shown to be superior over ANOVA based estimator (Huber, 1993; Searle et al. 1992). Like GAREML (Huber, 1993), this procedure can not only simultaneously perform variance component estimation and BLUP analysis of random genetic effects (e.g. GCA and SCA), but also provide flexibilities in data analysis (see chapter 1).

After variance components of random effects are estimated, the genetic variance and other parameters can be derived (using the formulae in Method section of Chapter 3). These parameters include: additive and non-additive variance, narrow sense individual heritability, broad sense heritability, half-sib family mean heritability, full-sib family mean heritability (narrow sense), full-sib family mean heritability (broad sense), narrow sense within full-sib family heritability, broad sense within full-sib family heritability.

To distinguish different selection methods that use the corresponding appropriate heritability in gain prediction, the following notation was used in this paper:

Table 4-2. Notation for selection methods and relevant heritability for calculating genetic gains used in this study

Notation	Selection method	Heritability
IND_A	Mass selection	Narrow sense individual heritability
IND_G	Clonal selection	Broad sense heritability
HS	Selection on half-sib family means	Half-sib family mean heritability
FS_A	Selection on mid-parent values	Full-sib family mean heritability (narrow sense)
FS_G	Selection of best full-sib families (Mass production)	Full-sib family mean heritability (broad sense)
FS+WFS_A	Selection of best individuals within best full-sib family	Both full-sib family heritability and within full-sib family heritability (narrow sense)
FS+WFS_G	Clonal selection of best individuals within best full-sib family	Both full-sib family heritability and within full-sib family heritability (broad sense)

Using the same variance estimation procedure, the genetic correlations were estimated on the sum of the values of two variable X and Y and then using the following formula:

$$r_{Gxy} = \frac{\sigma_{gxgy}}{\sqrt{\sigma^2_{gx} \cdot \sigma^2_{gy}}} = \frac{\left[\sigma^2_{gx+gy} - \sigma^2_{gx} - \sigma^2_{gy}\right]/2}{\sqrt{\sigma^2_{gx} \cdot \sigma^2_{gy}}} \qquad (2)$$

Where

X, Y is the two traits of interest (e.g. height or volume at any age)

σ_{gxgy} is the genetic covariance between two traits,

σ^2_{gx+gy} is the variance estimate of the sum of two additive genetic values of two traits.

Since the data were generally well balanced in terms of crosses and missing plots and the estimation of σ_e^2 was not of particular interest, plot means were instead used in the analysis to obtain σ^2_{gx+gy}.

Genetic parameters estimated were then averaged across all test series in a test region. Since these tests were well-balanced and share the same mating design with equal number of parents and full-sib families, all estimates of a parameter were assumed to have the similar precision. Hence no weighting function was needed to obtain average parameter estimate (Hodge, 1992). The simple unweighted average was calculated.

Genetic gain

Genetic gain in trait Y resulted from indirect selection on trait X can be calculated using the following:

$$CG_{x.y} = i \cdot r_{Gxy} h_x h_y \sigma_y \qquad (3)$$

Where j = selection age

i = selection intensity

r_{Gxy} = genetic correlation between selected trait and age 8 volume

h_x= square root of heritability of selected trait X

h_y = square root of heritability of response trait Y

Trait Y was the response trait, i.e. tree volume at age 8. It is used as the selection goal since this was the latest-year measurement for most test series. Trait X can be height or volume at any earlier age.

Correlated responses in age-8 volume were calculated for indirect selection on height and volume at age 6, the recommended selection age at current breeding program (Li, 1996; Balocchi, 1992). Different selection strategies were considered: individual or mass selection, individual clonal selection, backward selection on half-sib family, forward selection on full-sib family mean based on additive or total genetic component, full-sib family plus within family selection based on additive or total genetic component. For the example illustrated in Fig 4-1, individual selection and within family selection, the selection intensity i was set to 2.665. For family selections, a selection intensity of 1.755 was used.

Selection Efficiency

Different selection methods at different ages were evaluated by comparing indirect selection at earlier ages for each trait with direct selection on volume at age 8. For individual selection methods, mass selection was used for direct selection, while for family selection methods, full-sib family selection based on additive genetic effect was used. Comparison was not made between family selection methods and individual selection methods since different selection intensities are usually used for these two methods. In combined selection, intensity was set to be 2.665 (select top 1% trees) for within family selection and 1.755 (select top 10% families) for family selection.

Selection efficiency was calculated as either gain per year or present value. Gain per year was calculated assuming 3 years from measurement to seed collection (Matheson et al. 1994). Genetic gain was divided by time until genetic gain was realized, i.e. selection age plus 3 years. The selection efficiency (SE1) was defined as the ratio of gain per year between indirect selection and direction selection.

$$SE_1 = \frac{CG_{j\cdot 8}/(j+3)}{G_8/(8+3)} = \frac{i \cdot r_{j\cdot 8} h_j h_8 \sigma_8 /(j+3)}{i \cdot h_8^2 \sigma_8 /11} = \frac{11 \cdot r_{j\cdot 8} h_j}{(j+3) \cdot h_8} \qquad (4)$$

Where j = selection age

i = selection intensity

$r_{j.8}$ = genetic correlation between selected trait and age 8 volume

h_j= square root of heritability of selected trait

h_8 = square root of heritability of age 8 volume

Present value utilizes the discount equation (Balocchi et al. 1993, Mckeand 1988) with interest rate (I) fixed at 8%. The only assumption was that t additional years beyond selection were needed to realize genetic gain. Selection efficiency based on present value (SE2) was defined as the ratio of present value between indirect selection and direct selection.

$$SE_2 = \frac{PV_{j\cdot 8}}{PV_8} = \frac{CG_{j\cdot 8}/(1+I)^{j+t}}{G_8/(1+I)^{8+t}} = \frac{r_{j\cdot 8}h_j}{h_8}(1+I)^{8-j} \qquad (5)$$

where symbols were defined as same as (3). Notice t was canceled out in the final expression in the equation.

Selection using BLUP estimates

Using the mixed model analytical procedure described above, the annual general combining ability (GCA) for each half-sib family and the annual specific combining ability (SCA) for each full-sib family were obtained by BLUP methodology for height or volume.

$$\hat{\mathbf{A}} = \mathbf{C'}\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (6)$$

Where $\hat{\mathbf{A}}$ is the vector of predicted genetic effect (e.g. GCA, SCA),

$\mathbf{C}$ is covariance matrix of $\mathbf{A}$ and $\mathbf{Y}$;

$\hat{\mathbf{V}}$ is the estimate variance matrix of $\mathbf{Y}$;

$\mathbf{Y}$ and $\mathbf{X}$ are defined the same as in model (2) and $\mathbf{X}\hat{\boldsymbol{\beta}}$ is BLUE solution of $\mathbf{X}\boldsymbol{\beta}$.

BLUPs of GCA were adjusted using checklot means (formula 6, more details are described in chapter 2). Full-sib breeding value (BV) was calculated by adding two parental GCA estimates and full-sib genetic value was obtained by adding SCA estimate to BV. BV and GV were also adjusted using the following formulas (7, 8).

$$GCA^* = GCA_m + GCA_f + .5^*(\text{Diallel mean} - \text{Checklot mean}) \qquad (7)$$

$$BV^* = GCA_m + GCA_f + (\text{diallel mean} - \text{checklot mean}) \qquad (8)$$

$$GV^* = GCA_m + GCA_f + SCA_{mf} + (\text{diallel mean} - \text{checklot mean}) \qquad (9)$$

Selection was then applied on different traits at different age. Based on the absolute genetic gains in age 8 volume, the relative efficiencies of different selection strategies (based on half-sib GCA, full-sib GV and BV) were evaluated.

Results

Genetic gains

Genetic gain estimates ranged from .06 to .57 and varied for different regions, traits and different selection methods (Fig 4-1). For each selection method in each region, selection on volume at age 6 yielded greater gain in year-8 volume than selection on height. Among test regions, Coastal population had the greatest correlated response in year-8 volume to selection, followed by Piedmont population and Northern population.

Regardless of traits and test regions, similar patterns were observed when comparing different selection methods within a test region per trait (Fig 4-1). Gain from within family selection alone was less than individual selection whether clonal breeding technique was used or not. Among three family selections, full-sib selection based on total genetic component had the greatest genetic gain, followed by mid-parent full-sib selection (narrow sense) and half-sib family selection. Selection among full-sib families had about 40% genetic gain over selection on half-sib families.

Comparison between individual or within family selection and family selection was misleading, because selection intensities were not the same. However for such different selection intensities, which favor individual selection (1% for individual and within family selection and 10% for family selection), the magnitude of genetic gain from family selection was still equivalent to that of individual selection. Both were greater than within family selection.

For all selection methods, additional gain can be achieved by capturing non-additive genetic component through mass-producing full-sib crosses (full-sib selection) and vegetative propagation (FS+WFS_G). For individual selections, clonal selection increased gain 12-22% over mass selection. Selection of best full-sib families improved 10%-33% over selection of middle parent values. For family plus within family selection, this addition gain could be more than 40% (for Northern and Coastal population).

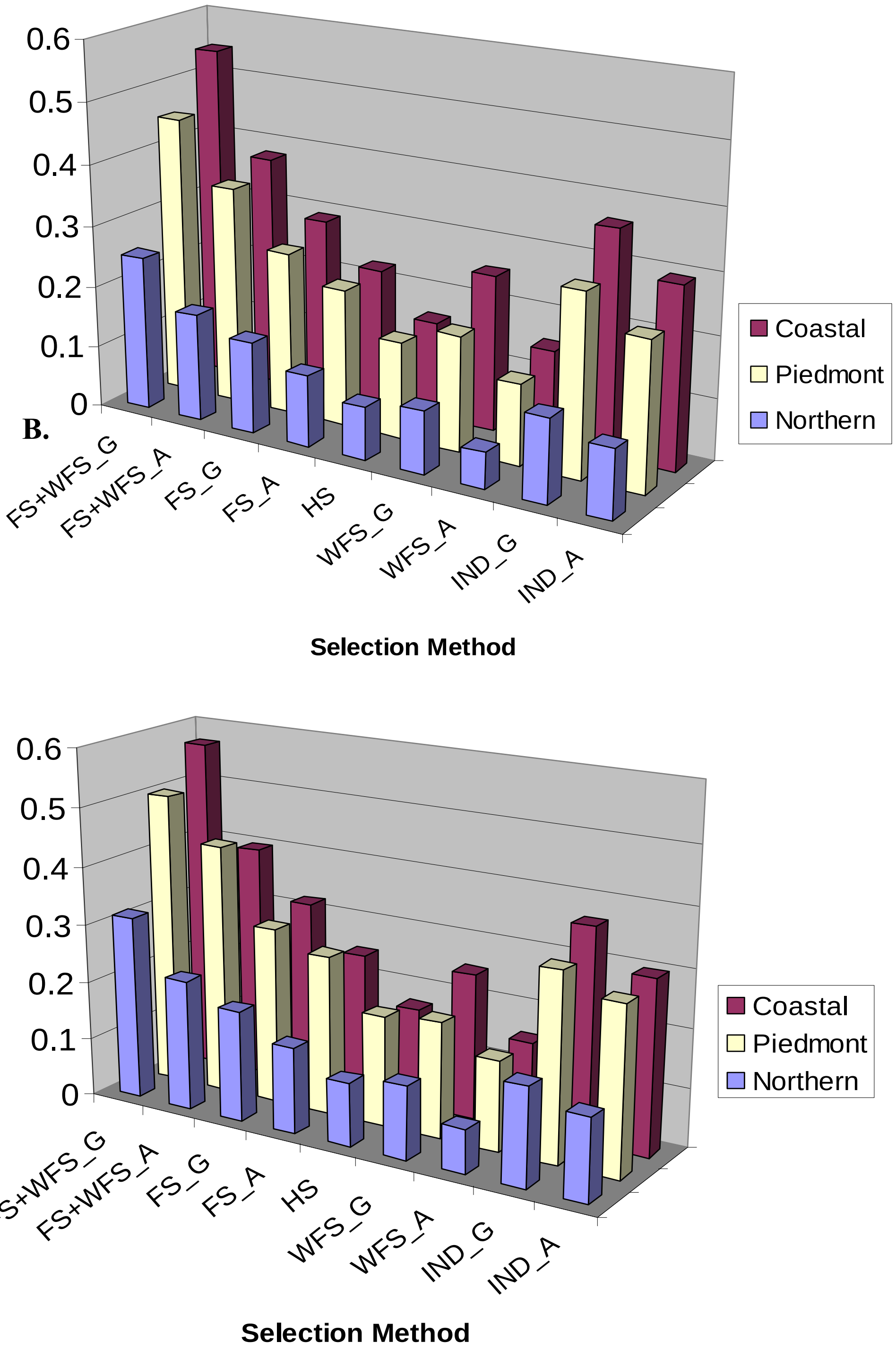


Figure 4-1. Correlated response of 8-yr volume from various selection methods on height (A) and volume (B) at age 6 in three test regions

Selection efficiency

The selection efficiency of various selection methods were plotted against age for both height and volume selection over three test regions (Appendix 4. Fig. 4-7 through Fig. 4-12). They indicated that compared with direct selection on year-8 volume, indirect selections at certain or most earlier ages were more efficient (i.e. Q>1) for most selection methods, except for individual selection for height.

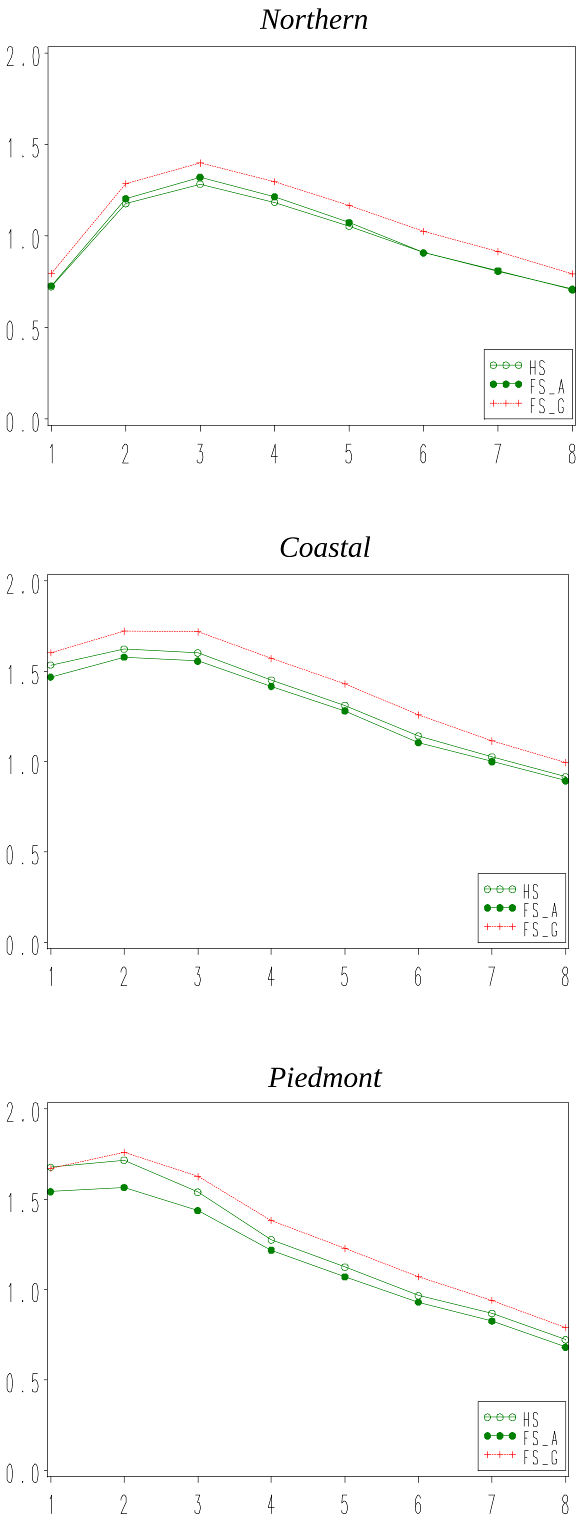
Results of selection efficiencies based on trait height for individual selection, family selection and combined family plus within family selection were summarized in Appendix 3. Fig. 4-7 through Fig. 4-9. Selection efficiency was calculated for selection based on both additive genetic variance and total genetic variance when appropriate.

Selection efficiency varied over time for different selection methods and test regions, with quite a large range, from .7 to 2. For all selection methods on height, the basic shape of selection efficiency over time had the same pattern with unique mode in somewhere between two extreme selection ages. Different test regions revealed different magnitudes of selection efficiency due to different time trends of genetic parameters (Fig 4-2).

In Northern region, all selection methods had the same optimal age year 3 for SE1. If SE2 was used as the criterion, the same optimal age was applicable to family selection, but the optimal age for SE2 delayed about one year for individual selection and combined selection.

The selection efficiency (both SE1 and SE2) was higher in Coastal region and Piedmont region than SE in Northern region (Fig 4-2). In addition, the peak value of SE appeared earlier for all selection methods in Piedmont region and for family selections in Coastal region (Fig 4-3). As a result, optimal age for selection on height could be very early. If gain per year was used as the criterion as in SE1, the optimal age was as early as age 2 for family selections in these two regions. For other selection methods, age 3 was the optimal age. If present value was instead used as criterion (SE2), age 3 was the most efficient for all selection methods in Piedmont region except half-sib family selection. Though in Coastal region optimal age for individual selection was delayed to age 5, age 3 was only slightly less efficient in SE2. Overall, age 3 was also a good choice for any selection method in Coastal region and Piedmont region.

For both SE1 and SE2, selection methods based on total genetic component had significant improvement in efficiency than those based solely on additive genetic component.

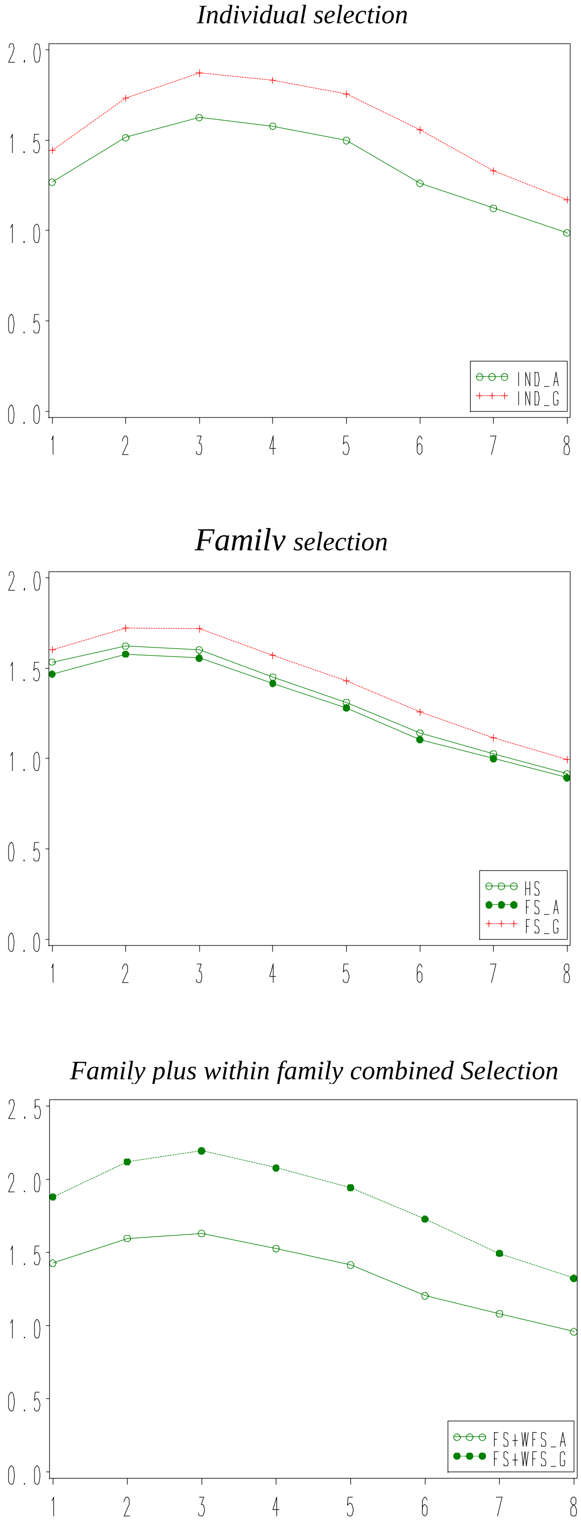


Northern

Coastal

Piedmont

Figure 4-2. Comparison of three test regions in family selection efficiencies of early height on 8-yr volume base on gain per year criterion (SE1) for three regions



Individual selection

Family selection

Family plus within family combined Selection

Figure 4-3. Comparison of different selection methods in selection efficiencies of early height on 8-yr volume for Coastal population, based on gain per year (SE1)

Results of selection efficiencies based on volume for individual selection, family selection, within family selection and combined family plus within family selection were illustrated in Appendix 3 Figure 4-10, 4-11, 4-12. Similar to height, selection efficiency was calculated for selection based on both additive genetic variance and total genetic variance when appropriate.

The range of SE varied from .7 to 2, with most SEs being nearly or exceeding one. Shape and magnitude of SE1 and SE2 varied from test region to test region (Fig 4-4). Although age 4 was the first evaluation age for volume, the change was quite evident for some SE curves for such a short period. For some selection methods, optimal age was achieved at the first measurement, i.e. age 4, indicating the true optimal age could even be earlier.

In Northern region, all selection methods showed that age 4 was the optimal age for GPY based selection efficiency (SE1). For individual selection and within family selection, age 5 was a lower point with increasing trend spreading towards both directions. Due to this irregular shape of the SE curve, the optimal ages for SE2 lagged behind about 2 years for individual selection and within family selection based on total genetic variance. For selection methods based on additive variance, optimal age was age 4. Even if age 5 or 6 was optimal, age 4 was only slightly less and should still be a good candidate year for selection. Both SE1 and SE2 were larger than 1 by fairly large margin at optimal ages, which indicated selection was more efficient than direct selection.

Within the measurement duration, age 4 was the optimal age for all selection methods in Coastal region regardless of selection criteria (Fig 4-5). After age 4, selection efficiency declined almost linearly. This pattern was mainly due to the facts that the age-age correlation for volume was very high in this test region and time discount function in the SE formula largely determined the shape of the curve. Based on time trend of selection efficiency, optimal age for selection on volume could be potentially at an earlier age that had been measured. In addition, the peak value of SE was very high being about 50%-100% more efficient than direct selection for SE1 and around 30%-90% more efficient than direct selection for SE2.

Selection efficiency for Piedmont region exhibited a bit more complicated pattern. Like Northern region and Coastal region, family selections had earlier optimal ages, with age 4 most of times. Individual and within family selection revealed different patterns for selection based on additive genetic variance and total genetic variance. SE based on total genetic variance had a lower point at age 5, while SE based on additive genetic variance peaked at age 6.

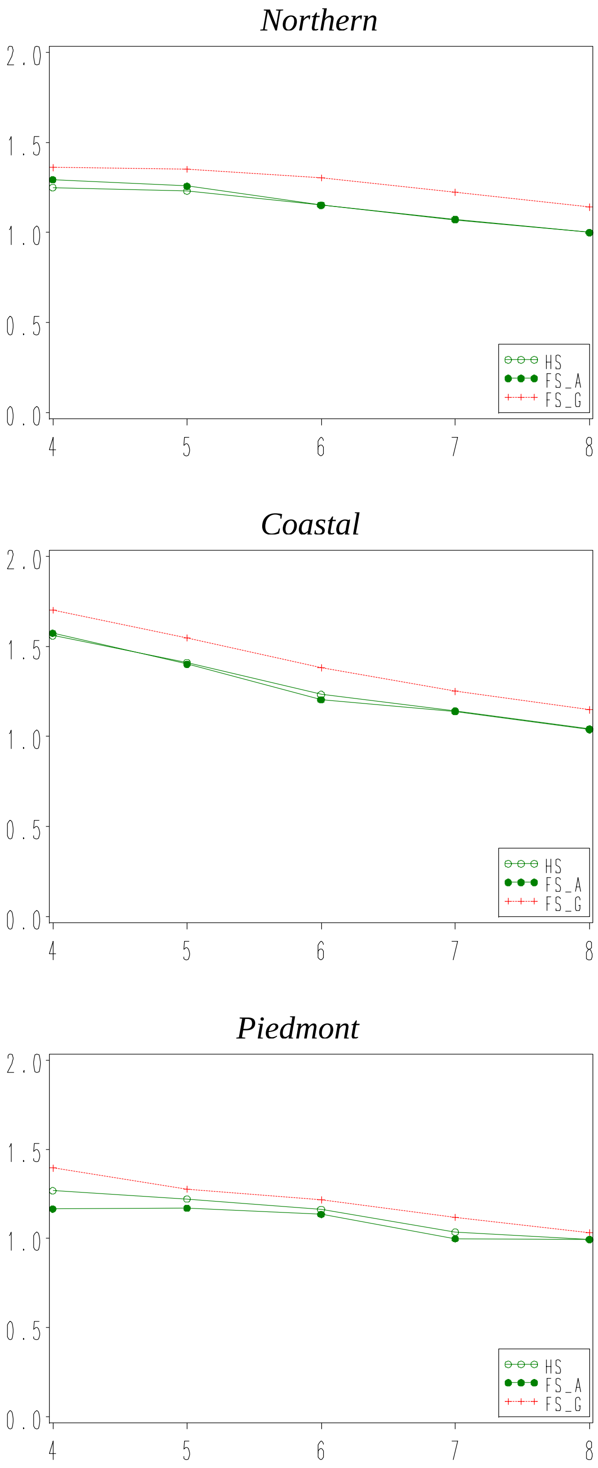
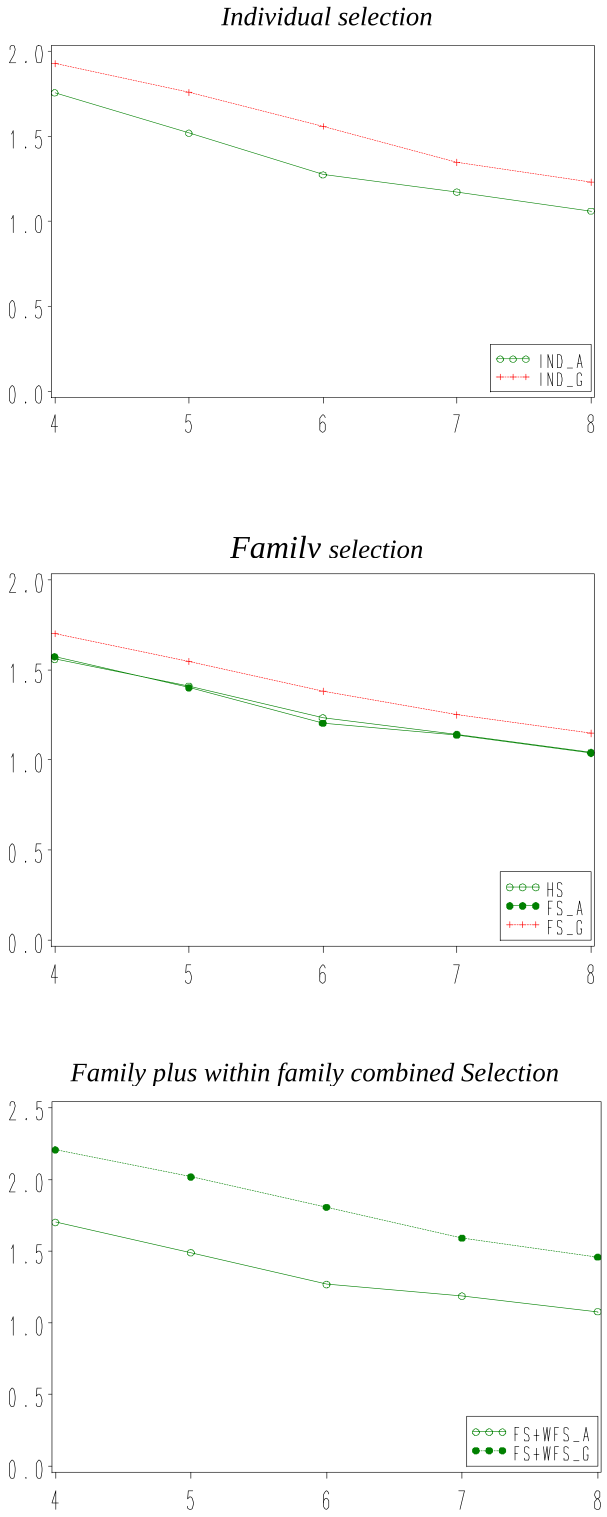


Figure 4-4. Comparison of three test regions in family selection efficiency of early volume on 8-yr volume base on gain per year criterion (SE1) for three regions

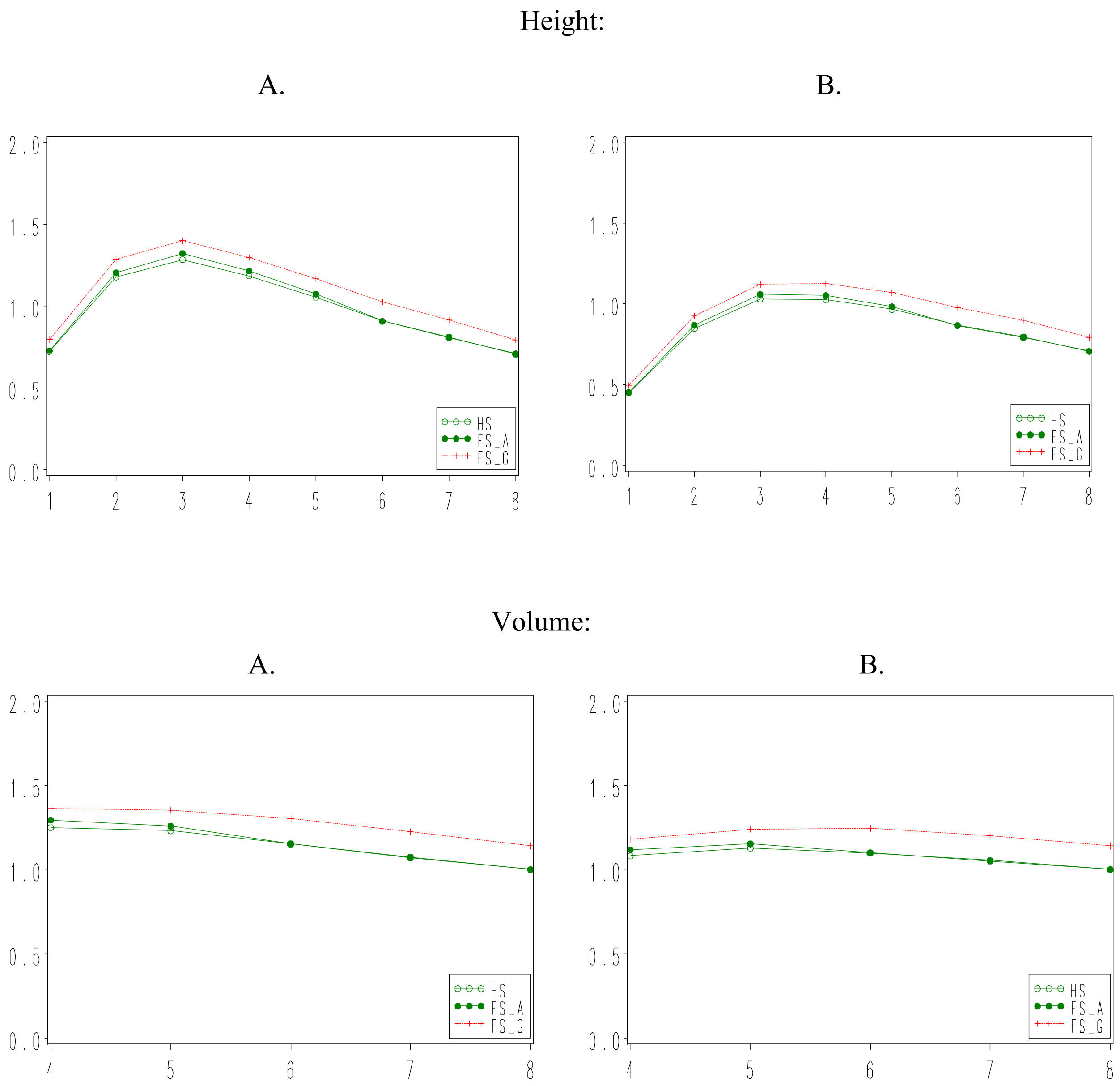


Individual selection

Familv selection

Family plus within family combined Selection

Figure 4-5. Comparison of different selection methods in selection efficiency of early volume on 8-yr volume for Coastal population, based on gain per year (SE1)


Height:

A.

B.

Volume:

A.

B.

Figure 4-6. Comparison of selection efficiency criteria, gain per year (A) vs present value (B) of early height or volume on 8-yr volume for Northern population

Selection Criteria: GPY vs PV

Both selection criteria considered time as a discount factor. In SE1, gain-per-year (GPY) was to maximize the genetic gain per unit of time in a breeding cycle (Li et al, 1996; Matheson et al, 1994). Present value (PV) in SE2 measured investment return. It emphasizes the reduction of value in the course of time by introducing accumulated interest rate power function into denominator (Balocchi et al, 1994; Mckeand and Bridgwater 1986). As an example for illustration, the comparison of time trend of selection efficiency for both criteria for height and volume in Northern region is shown in Fig 4-6.

For both selection criteria, one parameter needs to be subjectively determined. In case of GPY, additional time (t) for breeding cycle to complete may vary for various reasons such as breeding strategies and techniques utilized in a certain breeding program. Assuming 1, 3, 5 years needed to complete seed collection after measurement, selection efficiencies were summarized in table 4-3 and 4-4. Similarly for PV, additional time for breeding cycle (t) did not affect selection efficiency since it was canceled out in the ratio formula for SE2. But interest rate (I) can vary from time to time. While interest rate was fixed at 8% in our primary analysis, rate of 4% and 12% were also considered in the analysis. Results of optimal ages were summarized in table 4-3 and 4-4.

Table 4-3. Optimal age for various selection methods on height

Selection Method		Optimal Age					
		Northern region		Coastal region		Piedmont region	
		GPY*	PV*	GPY	PV	GPY	PV
Individual	IND*_A*	3 3 3	4 4 3	1 3 4	5 5 4	1 3 3	4 3 3
	IND_G*	3 3 4	5 4 4	1 3 4	5 5 4	1 3 3	4 3 3
Family	HS*	2 3 3	4 3 3	1 2 3	4 3 3	1 2 2	3 2 2
	FS*_A	2 3 3	4 3 3	1 2 3	4 3 3	1 2 2	3 3 2
	FS_G	2 3 3	4 3 3	1 2 3	4 3 3	1 2 2	3 3 2
Within Family	WFS*_A	3 3 3	4 4 3	2 3 4	5 5 4	1 3 3	4 3 3
	WFS*_G	3 3 4	5 5 4	2 3 4	5 5 4	1 3 3	5 3 3
Combined (family + within family)	FS+WFS_A	2 3 3	4 4 3	1 3 3	5 4 3	1 2 2	3 3 3
	FS+WFS_G	2 3 3	5 4 3	1 3 3	5 4 3	1 2 3	3 3 3

*Note: GPY: gain per year, optimal ages are listed in the order of t=1, 3, 5;

PV: present value, optimal ages are listed in the order of I=4%, 8%, 12%;

IND: individual selection, FS: full-sib family selection, HS: half-sib family selection

A: selection only on V_a, G: select on total genetic variance V_G

Table 4-4. Optimal age for various selection methods on volume

Selection Method		Optimal Age					
		Northern region		Coastal region		Piedmont region	
		GPY	PV	GPY	PV	GPY	PV
Individual	IND_A	4 4 4	7 4 4	4 4 4	4 4 4	6 6 6	8 6 6
	IND_G	4 4 6	7 6 4	4 4 4	5 4 4	4 4 7	7 7 6
Family	HS	4 4 5	6 5 5	4 4 4	4 4 4	4 4 4	6 5 4
	FS_A	4 4 4	5 5 4	4 4 4	4 4 4	4 4 5	6 6 5
	FS_G	4 4 5	7 5 5	4 4 4	4 4 4	4 4 4	6 4 4
Within Family	WFS_A	4 4 4	7 4 4	4 4 4	4 4 4	6 6 6	8 6 6
	WFS_G	4 4 6	7 7 4	4 4 4	5 4 4	4 4 7	7 7 4
Combined (family + within family)	FS+WFS_A	4 4 4	6 4 4	4 4 4	4 4 4	4 6 6	6 6 6
	FS+WFS_G	4 4 6	7 6 4	4 4 4	5 4 4	4 4 4	7 4 4

*Note: same notation as in table 2.

Selection Using BLUP

The results of BLUP selection on GCA of height and volume were listed in table 4-5 and the BLUP selection on full-sib genetic value (GV) and breeding value of both traits was illustrated in table 4-6 for height and in table 4-7 for volume. A range of selection intensities (50%, 25%, 10%, 5%) were selected to investigate its actual impact on gain based on year-8 volume BLUPs.

For a typical selection intensity i=1.775 (top 10% parents), the range of genetic gain estimated from BLUP selection of year-6 parental GCA was from .20 to .38 for height and from .42 to .44 for volume, which were higher than response from half-sib family selection based on heritability.

Gain differences among test region were similar to the results from usual genetic gain calculation using heritability estimates (refer to the first section of results). In each region, each selection method applied on volume usually produced greater gain in year-8 volume than selection on height. Among 3 test regions, selection in Coastal region had the greatest genetic gain, followed by Piedmont and Northern population. Due to the finite family size, the above general trends did not follow exactly for each test region or each trait, especially for higher selection intensity.

Over time genetic gain had a general increase trend, especially for the least aggressive selection intensity (i=.798, p=50%). But the increase over time was slow and not consistent. In fact, except for the first year all other ages had very competitive genetic gain as the final year for height. For volume, not much difference in genetic gain was observed for all ages.

As expected, genetic gain increased as increasing in selection intensity for selection on volume. This was also generally true for selection on height but the highest selection intensity did not guarantee the greatest gain. For example, for age 3 and 4 in Coastal region, selection of 10% had slightly less gain than that of 25%. The decrease in genetic gain occurred more frequently when selection percentage was moved from 10% up to 5%. Finally, for both traits in Coastal region, there was not much difference between the gains from two highest selection intensities (for 10% and 5%).

Table 4-5. Estimated genetic gain in 8-year volume from selection on BLUP of GCA of early height and volume

		Selection on height					Selection on volume				
Test region	Age	# of family	50%	25%	10%	5%	# of family	50%	25%	10%	5%
Northern	1	36	0.1381	0.1844	0.1757	0.2087					
	2	48	0.1258	0.1719	0.2687	0.3675					
	3	48	0.1267	0.1629	0.1745	0.1487					
	4	48	0.1513	0.1975	0.2482	0.3675	48	0.1565	0.1996	0.2566	0.2864
	5	48	0.1536	0.1956	0.2482	0.3675	48	0.1590	0.2128	0.2713	0.3675
	6	48	0.1500	0.1926	0.2015	0.1487	48	0.1568	0.2137	0.2713	0.3675
	7	48	0.1500	0.1857	0.2015	0.1487	48	0.1678	0.2141	0.2728	0.3675
	8	48	0.1608	0.1956	0.2415	0.2962	48	0.1682	0.2157	0.2793	0.3675
Coastal	1	84	0.2112	0.2471	0.3724	0.3965					
	2	84	0.2014	0.2627	0.3766	0.3237					
	3	84	0.2402	0.2582	0.2438	0.2239					
	4	84	0.2436	0.3075	0.3031	0.3425	84	0.2466	0.3184	0.3952	0.4253
	5	84	0.2576	0.3089	0.3575	0.4250	84	0.2681	0.3264	0.4282	0.4703
	6	84	0.2570	0.2712	0.2902	0.3270	84	0.2649	0.3313	0.4282	0.4759
	7	84	0.2512	0.2996	0.3198	0.3969	84	0.2607	0.3224	0.4316	0.4868
	8	84	0.2581	0.2963	0.3171	0.3162	84	0.2727	0.3402	0.4407	0.4917
Piedmont	1	60	0.1380	0.1645	0.1718	0.0857					
	2	72	0.1851	0.2576	0.3574	0.4370					
	3	72	0.2288	0.2971	0.3600	0.4370					
	4	60	0.2334	0.2393	0.2729	0.2816	48	0.1748	0.2155	0.2377	0.3221
	5	72	0.2296	0.2763	0.3055	0.3826	60	0.1920	0.2361	0.3011	0.3213
	6	72	0.2293	0.3079	0.3833	0.4189	72	0.2435	0.3147	0.4196	0.4566
	7	60	0.1888	0.2147	0.2645	0.3153	60	0.1975	0.2502	0.3133	0.3403
	8	72	0.2406	0.3020	0.3833	0.4189	72	0.2554	0.3253	0.4196	0.4588

When BLUP selection was carried on year-6 height with a typical selection intensity i=1.775, the range of genetic gain was from .33 to .71 for full-sib GV and from .30 to .70 for full-sib BV. For selection on volume, genetic gain was from .54 to .75 for full-sib GV and from .52 to .75 for full-sib BV. Again these figures were higher than those calculated from heritability estimates.

Like gain prediction for GCA, gain differences among test region were also similar to those calculated using heritability estimates (refer to the first section of results). For each regional population, selection on volume usually yielded greater gain in year-8 volume than selection on height. Selection in Coastal region had the greatest genetic gain among 3 test regions, followed by Piedmont population and Northern population. These general trends did not follow strictly for each population or each trait due to the finite family size, especially when more aggressive selection intensity was chosen.

The genetic gain generally increased over time, most evidently for the lowest selection intensity (i=.798, p=50%). However the increase trend was slow and inconsistent. Except for the first year selection on other ages yielded very competitive genetic gains as the final year for height. For volume, genetic gains were around the same range for all ages, except for some abrupt changes in Piedmont population. These changes may be caused by less available full-sib families at certain ages since the common checklot of that region was missing in one test series.

Since family size was much larger for full-sib family selection than for GCA selection, the increase in genetic gain for higher selection intensity was more consistent for both traits. Occasionally, a small decrease in genetic gain was still possible when selection percentage was moved from 10% up to 5%.

Table 4-6. Estimated genetic gain in 8-yr volume from BLUP selection of full-sib GV or BV of early height

			Selection based on full-sib GV (GCA_m+GCA_f+SCA)				Selection based on full-sib BV (GCA_m+GCA_f)			
Test region	AGE	# of family	50%	25%	10%	5%	50%	25%	10%	5%
Northern	1	90	0.2851	0.3067	0.4370	0.5352	0.2781	0.3214	0.4321	0.3945
	2	120	0.2653	0.3137	0.4680	0.5352	0.2584	0.3101	0.4713	0.5341
	3	120	0.2467	0.2558	0.3108	0.3122	0.2399	0.2358	0.2988	0.3343
	4	120	0.3030	0.3726	0.4312	0.5352	0.3025	0.3633	0.4369	0.5341
	5	120	0.3144	0.3956	0.4649	0.6075	0.3158	0.3945	0.5107	0.5341
	6	120	0.2859	0.3490	0.3302	0.3966	0.2771	0.3869	0.3023	0.3598
	7	120	0.3094	0.3652	0.3701	0.3966	0.3001	0.3811	0.3065	0.3644
	8	120	0.3167	0.3828	0.4281	0.4749	0.3204	0.3799	0.4259	0.4366
Coastal	1	208	0.4290	0.4937	0.6571	0.6618	0.4220	0.4846	0.6510	0.6618
	2	208	0.3979	0.5420	0.6593	0.7180	0.3895	0.5223	0.6927	0.7633
	3	208	0.4423	0.4520	0.4477	0.5404	0.4372	0.4477	0.4113	0.4903
	4	208	0.4688	0.5427	0.6225	0.6413	0.4579	0.5164	0.5894	0.5971
	5	208	0.4924	0.5943	0.6883	0.7132	0.4820	0.5902	0.6785	0.6811
	6	208	0.4717	0.5316	0.5444	0.5663	0.4705	0.5182	0.5248	0.5347
	7	208	0.4917	0.5611	0.6078	0.5636	0.4697	0.5307	0.5307	0.5484
	8	208	0.4948	0.5539	0.6062	0.6016	0.4855	0.5344	0.5711	0.5237
Piedmont	1	144	0.2557	0.3174	0.3272	0.2456	0.2613	0.3184	0.3286	0.2483
	2	175	0.3579	0.5082	0.6814	0.7387	0.3432	0.5115	0.6817	0.7050
	3	175	0.4443	0.5665	0.7037	0.7848	0.4447	0.5657	0.6978	0.7757
	4	144	0.4479	0.4510	0.4804	0.5424	0.4468	0.4523	0.4752	0.5408
	5	172	0.4410	0.5135	0.5521	0.5146	0.4381	0.5028	0.5446	0.5053
	6	174	0.4454	0.5741	0.7087	0.7786	0.4442	0.5640	0.7014	0.7756
	7	144	0.3484	0.4066	0.4850	0.5444	0.3483	0.3902	0.4621	0.5264
	8	174	0.4613	0.5732	0.7138	0.7786	0.4620	0.5740	0.7079	0.7780

Table 4-7. Estimated genetic gain in 8-yr volume from BLUP selection of full-sib GV or BV of early height

			Selection based on full-sib genetic value (GCA_m+GCA_f+SCA)				Selection based on full-sib genetic value (GCA_m+GCA_f)			
Test region	AGE	# of family	50%	25%	10%	5%	50%	25%	10%	5%
Northern	4	120	0.3016	0.3906	0.4992	0.6192	0.2970	0.3835	0.5026	0.5978
	5	120	0.3136	0.4055	0.5269	0.6369	0.3098	0.4003	0.5234	0.6220
	6	120	0.3161	0.4105	0.5351	0.6369	0.3116	0.4013	0.5214	0.6220
	7	120	0.3226	0.4190	0.5413	0.6369	0.3179	0.4082	0.5258	0.6220
	8	120	0.3238	0.4208	0.5485	0.6369	0.3188	0.4110	0.5389	0.6220
Coastal	4	206	0.4785	0.5590	0.6943	0.8197	0.4661	0.5434	0.6904	0.8194
	5	207	0.5140	0.6009	0.7297	0.8551	0.5009	0.5788	0.7167	0.8513
	6	207	0.5121	0.6140	0.7378	0.8668	0.4992	0.5904	0.7054	0.8464
	7	207	0.4970	0.6167	0.7312	0.8627	0.4832	0.5928	0.6943	0.8549
	8	208	0.5197	0.6343	0.7554	0.8669	0.5079	0.6046	0.7220	0.8552
Piedmont	4	142	0.3472	0.4091	0.4747	0.5144	0.3256	0.4113	0.4776	0.5378
	5	143	0.3663	0.4373	0.5158	0.5758	0.3642	0.4338	0.5107	0.5550
	6	174	0.4654	0.5945	0.7474	0.8302	0.4617	0.5882	0.7455	0.8211
	7	138	0.3763	0.4431	0.5378	0.5893	0.3725	0.4401	0.5228	0.5731
	8	174	0.4824	0.6089	0.7554	0.8576	0.4791	0.6018	0.7460	0.8439

Discussion

Comparison among genetic gains from various selection methods at current selection age (year 6) revealed that selection on volume yielded more gain than selection on height. This was primarily because of the higher genetic correlation of volume with age-8 volume and comparable heritability when compared with height (see chapter 3). This confirms the earlier results (Annual report of NCSU-ICTIP, 1999; Li, 1996) and that volume should be considered in early selection of loblolly pine in order to achieve the higher genetic gain.

Among test regions, Coastal population had the greatest correlated response in year-8 volume to selection, followed by Piedmont population and Northern population. Such a difference in gain prediction is expected because of their different growth rate. The Atlantic Coastal source is noted for its faster growth than other two sources. The Northern population is cold-hardy but grows relatively slow. The Piedmont source is also cold-hardy but is intermediate in growth pattern (NCSU-ICTIP,1999).

Selection among full-sib families had about 40% more genetic gain over selection on half-sib families. Family plus within family selection based on total genetic component can capture the most genetic gain. In the combined selection, family selection contributed more to the total genetic gain than within family selection. Similar results were reported for height selection in unimproved population of loblolly pine (Balocchi, 1990).

Non-additive genetic component was found to be very important and even exceed additive component at some time point in early growth stage of loblolly pine for growth trait (Balocchi, 1990; McKeand, 1986; Foster & Bridgwater, 1986). In the present study of early diallel tests, dominance component accounted for 20% to 40% of total genetic variance (chapter 3). Though not as high as unimproved population, it is still important for tree breeder to consider and capture this additional genetic gain through mass controlled pollination or vegetative propagation techniques such as rooted cutting. As shown in this study, for all selection methods, additional gain can be achieved by capturing partial non-additive genetic component through mass production of best full-sib families, and all of non-additive component through vegetative propagation in clonal selection, or through vegetative propagation of the best individual trees in the best full-sib family. In the example of selection at age 6, clonal selection increases as high as 22% gain over mass selection. Selection of best full-sib families can improve up to 33% over selection of middle parent values. Selection of the best individuals within the top full-sib family through vegetative propagation tops selection through usual breeding cycle by as much as 40% (for Northern and Coastal population).

Selection efficiency is a comprehensive statistics that combines the information of genetic parameters and discount factor of time. These genetic parameters are genetic correlation r_{j8} and heritability of both indirect and direct selection ages (h^2_j and h^2_8). The time trend of any of these three genetic parameters will affect the result of selection efficiency. Since heritability is usually of the same magnitude, genetic correlation has more influence on selection efficiency.

If selection is based on trait height of Northern population, the optimal selection age for all selection methods is age 3 based on SE1 criterion and selection is more efficient than direct selection at age 3. Based on SE2 criterion, about one year delay is observed in optimal age for individual selection and combined selection. But the optimal age does not change for family selection. SE2 also indicates that based on present value indirect selection on height can be nearly as efficient as direct selection on volume.

The higher selection efficiency on height for both SE1 and SE2 in Coastal region and Piedmont region than that in Northern region is mainly due to the higher genetic correlation between height and age 8 volume in these two regions (see chapter 3). Very early optimal ages for selection on height (age 2-3) imply that selection on height at very early growth stage has the potential to increase both gain per year and present value. Finally, additional improvement in efficiency is achieved for selection methods based on total genetic component have over those based solely on additive genetic component.

Results from selection efficiency study of Balocchi (1992) showed that optimal family selection age based on total genetic component was 2-3 year earlier than family selection based solely on additive component. In contrast, we found that selecting additional non-additive component does not have a significant effect on optimal age for selection on trait height. There are very little differences in optimal ages between these two strategies for selection on height (table 4-3). This is also true for volume selection in Northern region and Coastal region. For selection on volume in Piedmont region however, capturing non-additive genetic component does have advantage of earlier optimal selection age for family and combined family plus within family selection besides higher genetic gain.

Compared with gain prediction using population genetic parameter estimates, BLUP produce more desirable gain prediction in terms of various statistical properties (Borralho, 1995; Huber, 1992; White T. L. & Hodge, 1988, 1989). In this study, through mixed analytical procedure, BLUP prediction of genetic effects parental GCA and full-sib SCA were calculated for all ages. Genetic gains estimated from BLUP selection of year-6 parental GCA, full-sib GV and BV are higher than those from family selections based on heritability at the same selection intensity. It must be noted that genetic gain estimated from global BLUP measures the improvement over unimproved checklots, while gain prediction based on heritability predict genetic gain over current 2nd generation breeding population. In addition, checklot mean adjustment used for global BLUP puts all families from multiple test series on the same baseline for selection. The improvement in accuracy of prediction has been supported by computer simulation study (see chapter 2). Hence, the greater gain is expected for adjusted BLUP estimates.

Other than the higher gain prediction, similar results as the usual genetic gain based on heritability were found for BLUP selection. First, selection on volume usually produces greater gain in year-8 volume than selection on height due to the underlining difference in genetic parameter structure. Secondly, the same order of the genetic gain for 3 test regions was found (from the largest to the smallest): Coastal population, Piedmont population and Northern population. Finally, additional gain can be captured by selecting on full-sib family genetic value, which includes the SCA effect. This again reminds breeders of the importance of capturing non-additive genetic components in developing breeding strategy.

High selection intensity is not recommended for family selection because of the relatively small sample size for family selection. The consequence of inbreeding and reduced population size for future generation may cause serious problems (Falconer, 1989). In addition, the selection intensity study on BLUP selection showed another reason why a very high selection intensity is not necessarily a good solution to improve gain. Genetic gain from BLUP selection increased with the increasing selection intensity but the highest selection intensity at earlier ages did not always yield the greatest gain for volume at age 8. There is a possibility that an aggressive 5% selection percentage may cause a reduction in genetic gain from selection of 10%. However as the number of families increases, this reduction become less serious.

Conclusion

The genetic gains predicted for various selection methods at age 6 revealed that selection on volume yields more gain than selection on height. Family plus within family selection based is the most effective to achieve genetic gain for early selection on both height and volume. Additional gain can be achieved by capturing non-additive genetic component through mass-producing full-sib crosses and vegetative propagation.

The analysis of selection efficiency showed earlier selection could be more efficient than direct selection on volume of age 8. Family selection can be performed as early as age 2 or 3 for trait height and at age 4 for DBH or volume. Family plus within family has its optimal selection age at age 3 or age 4. The results from BLUP selection indicated similar results as the usual genetic gain based on heritability except that genetic gains estimated from BLUP selection are higher than those from family selections based on heritability.

The implications of this study are: 1) volume should be considered in early selection of loblolly pine in order to achieve maximum genetic gain; 2) selection at earlier age (e.g. 3 or 4) should be considered to achieve the maximum gain per time unit in the loblolly pine breeding program; 3) different results from different test regions with regarding to the optimal age and selection efficiency level indicate separate consideration needs to be applied to selection strategy within each test region.

Reference

[1] Balocchi, C.E. et al. Age trends in genetic parameters for tree height in a nonselected popultaion of loblolly pine. *For. Sci.* 39:231-251. 1993.

[2] Balocchi, C.E. Age trends of genetic parameters and selection efficiency for loblolly pine (*Pinus Taeda* L.). Ph. D. dissertation. North Carolina State University. 1990.

[3] Borralho, N.M.G. 1995. The impact of individual tree mixed models (BLUP) in tree breeding strategies. In: Eucalypt plantations: improving fibre yield and quality. Proceedings of CRCTHF-IUFRO Conference. Hobart, Australia. p141-145.

[4] Falconer, D.S. Introduction to quantitative genetics. Longman & Co., New York, NY. 1989.

[5] Foster, G.S. Trends in genetic parameters with stand development and their influence on early selection for volume growth in loblolly pine. Forest Sci. 32: 4. 1986.

[6] Gwaze, D.P., Woolliams, J. A. and Kanowsk. Genetic parameters for height and stem straightness in *Pinus Taeda* Linnaeus in Zimbabwe. *For. Genet.* 4(3):159-169. 1997.

[7] Hodge, G.R. and White, T.L. Genetic parameter estimates for growth traits at different ages in slash pine and some implications for breeding. *Silvae Genetica* 41(4-5):252-262 1992.

[8] Huber, D.A. Optimal mating designs and optimal tehcniques for analysis of quantitative traits in forest genetics. Ph. D. dissertation. University of Florida. 1993

[9] Lantz, C.W., Kraus, J.F. A guide to southern pine seed sources. Southeastern Forest Experiment Station, Asheville, N.C. 1987.

[10] Li B., McKeand, S.E., Hatcher, A.V., and Weir, R.J. 1997. Genetic gains of second generation selections from the NCSU-Industry Cooperative Tree Improvement Program. Pro. 24th South. For. Tree Impr. Conf., p. 234-238. Orlando, Florida, June 9-12, 1997.

[11] Li, Bailian, McKeand, S.E., Weir, R.J. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the south-eastern US. p. 164-68. In: Tree improvement for sustainable tropical forestry. Proceedings of QFRI-IUFRO Conference. Caloundra, Queensland, Australia. 1996.

[12] Littell, R.C. et al. SAS system for mixed models. SAS Institute Inc. Cary, NC. 1996

[13] Mckeand, S.E., and Svensson, J.. Sustainable management of genetic resources. *Journal of Forestry* 94(3). 1997.

[14] Mckeand, S.E. Optimum age for family selection for growth in genetic tests of loblolly pine. *For. Sci.* 34:400-411. 1988.

[15] Matheson, A.C., Spencer, D. J. and Magnussen, D. Optimum age for selectio in Pinus radiata using basal area under bark for age:age correlations. *Silvae Genetica* 43: 352-257. 1994.

[16] SAS Institute Inc. SAS/STAT Software: Changes and enhancements (through release 6.11). Cary, NC. 1996.

[17] Searle, S.R., Casella, G., and Mcculloch, C.E. Variance components. John Wiley & Sons Inc., New York, 501p. 1992.

[18] van Buijtenen J. P. and Burdon R. D. Expected efficiencies of mating designs for advanced generation selection. *Can. J. For. Res.* 20:1648-1663. 1990.

[19] van Buijtenen, J. P. and Bridgwater, F. Mating and genetic test designs. In: Advanced generation breeding of forest trees. Southern Coop. Series Bull. 309. Louisiana Agr. Exp. Stn. Baton Rouge, LA. pp5-10.

[20] White, T.L. and Hodge, G. R. Best linear prediction of breeding values in a forest tree improvement program. *Theor. Appl. Genet.* 76:719-727. 1988.

[21] White, T.L. and Hodge, G. R. Predicting breeding values with applications in forest tree improvement. Kluwer Academic Pub., Dordrecht, The Netherlands. 367pp. 1989.

[22] Williams, C.G. and Megraw, R. A. Juvenile-mature relationships for wood density in *Pinus Taeda*. *Can. J. For. Res.* 24: 714-722. 1994.

[23] Yanchuk, A. D. General and specific combining ability from disconnected partial diallels of coastal douglas-fir. *Silvae Genetica* 45(1):37-45 (1996).

[24] Zobel, B.J. and John Talbert, J.T. Applied forest tree improvement. John Wiley and Sons. New York. NY. 1984.

Appendix 1: Outlines for analysis of diallel tests via SAS PROC IML and PROC MIXED

We start from the data set DIALL in the SAS system. In this example, it should at least have six variables BLOCK, FEMALE, MALE, CROSS, TREE and HEIGHT, with total N observations.

Step one is to read FEMALE, MALE variables into a matrix FE (N×2) in IML procedure and form a parent vector P (6×1) with parent names as its elements. Then a dummy variable matrix D (N×6) with "1" in corresponding columns for two parents in each row is constructed by comparing each row of FE with all parents in P. If the data is balanced, the matrices FE, P and D look like:

$$\mathrm{FE} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{2} \\ \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{6} \\ \boldsymbol{2} & \boldsymbol{3} \\ \vdots & \vdots \\ \boldsymbol{5} & \boldsymbol{6} \end{bmatrix}_{1500\times 2} \qquad \mathrm{P} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}_{6\times 1} \qquad \mathrm{D} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1} \\ \boldsymbol{0} & \boldsymbol{1} & \boldsymbol{1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1} & \boldsymbol{1} \end{bmatrix}_{1500\times 6}$$

Where 0, 1, 2, … , 6 are 100×1 vectors with identical elements (e.g. 1= [1, …, 1]′100×1).

Step two is to output a data set DUMM from dummy variable matrix D in IML and merge data set DUMM with the original data set DIALL.

Step three is to analyze the data with MIXED procedure with fixed effect TEST in MODEL statement and dummy variables P1-P6 and other random effects in multiple RANDOM statements. The option "TYPE=TOEP(1)" must be used to constrain the same variance over 6 dummy variables and over interactions P1*TEST,…, P6*TEST. The default variance estimation method is REML. Without missing parent and cross, BLUP of GCA, SCA and GLS solutions of fixed effects can be calculated using ESTIMATE statements. Otherwise it is more convenient to use MAKE statements to output solutions of random effects and fixed effects.

Finally, if individual tree breeding value is of interest, we can require the MIXED procedure to output variance components estimates, inverse of variance matrix of Y (V-1) and BLUE solution of Xβ ($\mathbf{X}\hat{\boldsymbol{\beta}}$) into three data sets. Then we read them into matrices in PROC IML and use formula (8) to calculate within family breeding value. The job is finished when $\hat{\mathbf{A}}_{\mathbf{w}}$ is added back to $\hat{\mathbf{G}}$.

Appendix 2: SAS code for half-diallel analysis

```
/*------------------------FILE: PAPER1.SAS-----------------------------
|                                                                     |
|                           BY BIN XIANG                              |
|                                                                     |
|                     FOR ILLUSTRATION PURPOSE                        |
|                                                                     |
| The following code is to retrieve the Diallel Test data (individual |
| tree file), create dummy variables for each parent in IML, analyze  |
| the data using MIXED PROCEDURE and calculate individual within      |
| family breeding value.                                              |
|                                                                     |
| The data structure is as described in METHOD section: RCBD with 5   |
| tests, 4 reps(blocks), 6 parents, 5 trees.                          |
|                                                                     |
---------------------------------------------------------------------*/

PROC IML;
  use DIALL;
  read all var {FEMALE MALE} into FE;
  N=NROW(FE);

  P=SHAPE('00000',6,1);
  P[1,1]=FE[1,1]; P[2,1]=FE[1,2]; K=2;
  DO I=1 TO N;F=0;M=0;
    DO J=1 TO K;
      IF FE[I,1]=P[J,1] THEN F=1;
      IF FE[I,2]=P[J,1] THEN M=1;
    END;
    IF F=0 THEN DO;
      K=K+1; P[K,1]=FE[I,1];
    END;
    IF M=0 THEN DO;
      K=K+1; P[K,1]=FE[I,2];
    END;
  END;

  D=SHAPE(0,N,6);
  DO I=1 TO N;
    DO J=1 TO 6;
      IF FE[I,1]=P[J,1] | FE[I,2]=P[J,1] THEN D[I,J]=1;
    END;
  END;
create DUMM from D [colname=s];
append from D [colname={'P1' 'P2' 'P3' 'P4' 'P5' 'P6'}];
QUIT;

DATA DIALL;
MERGE DIALL DUMM;
RUN;
```

```
PROC MIXED DATA=DIALL COVTEST; CLASS CROSS BLOCK;
     MODEL HEIGH=TEST/PM;
     RANDOM BLOCK(TEST);
     RANDOM P1-P6/TYPE=TOEP(1) VI;
     RANDOM CROSS;
     RANDOM P1*TEST P2*TEST P3*TEST P4*TEST P5*TEST P6*TEST/TOEP(1);
     RANDOM CROSS*TEST CROSS*BLOCK(TEST);
     ESTIMATE 'GCA1' | P1 1;
       ...  ...
     ESTIMATE 'GCA6' | P6 1;
     ESTIMATE 'SCA1' | CROSS 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
       ...  ...
     ESTIMATE 'SCA15' | CROSS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1;
     ESTIMATE 'TEST1' INT 1 BLOCK 1 0 0 0 0;
       ...  ...
     ESTIMATE 'TEST5' INT 1 BLOCK 0 0 0 0 1;
     MAKE 'VI' OUT=VI NOPRINT;
     MAKE 'PREDMEANS' OUT=PM NOPRINT;
     MAKE 'COVPARMS' OUT=COV;
RUN;

PROC IML;
  USE COV;
  READ ALL VAR {EST} INTO COV;
  CLOSE COV;
  USE PM;
  READ ALL VAR {_RESID_} INTO Y_XB;
  CLOSE PM;
  USE VI;
  READ ALL INTO VI;
  CLOSE VI;

  N=NROW(VI); K=NCOL(VI); VI=VI[,2:K];
  AW=2*COV[2]*VI*Y_XB;
  create AW from AW [colname='AW'];
  append from AW [colname='AW'];

QUIT;
```

Appendix 3: λ_{min} Values for Best Variance Predictor

PN*	e*	Mean*	S. D.	Minimum	Maximum
1	4	.4711	.1000	.0921	.6969
2	4	.5707	.0837	.2170	.7432
3	4	.4366	.1089	.0544	.6843
4	4	.4691	.1038	.1146	.7061
5	4	.5178	.0944	.0930	.7408
6	4	.6295	.0763	.2599	.7978
7	4	.4200	.1072	.0882	.6610
8	4	.5171	.0948	.1518	.7167
1	8	.4769	.0727	.1746	.6422
2	8	.5747	.0608	.3086	.7200
3	8	.4248	.0837	.1536	.6219
4	8	.4778	.0760	.2036	.6673
5	8	.5177	.0673	.2270	.6900
6	8	.6327	.0487	.4421	.7331
7	8	.4272	.0758	.1371	.6227
8	8	.5207	.0701	.2730	.6895

* PN — labels for parameter structure, see Table 2.2.1
e — the number of test series in a test region
mean — averaged over 1000 simulation runs

Appendix 4: Graphs for selection efficiency study

Figure 4-7. Selection efficiency of early height on 8-yr volume for Northern population, based on gain per year (SE1) or present value (SE2)

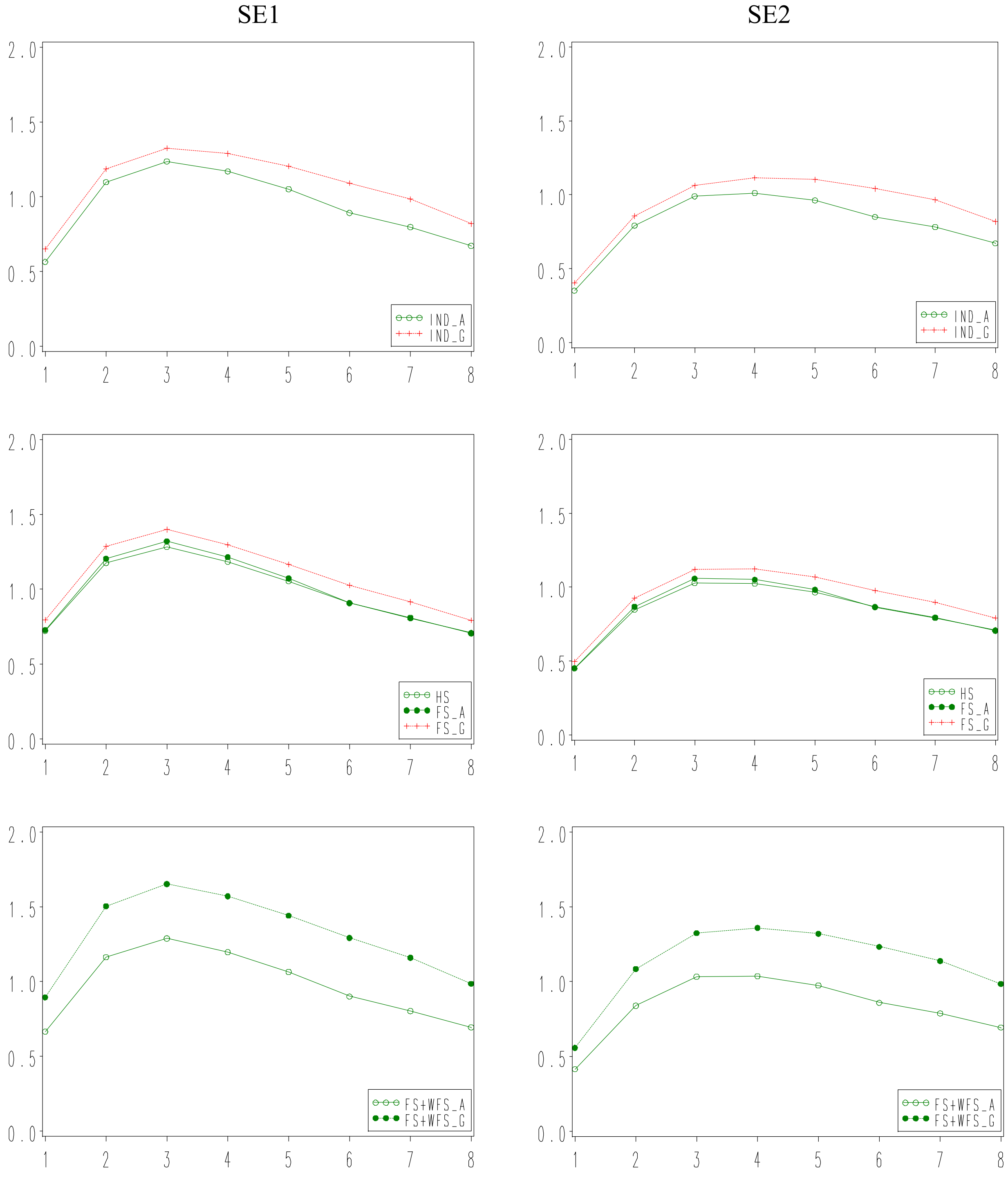


Figure 4-8. Selection efficiency of early height on 8-yr volume for Coastal population, based on gain per year (SE1) or present value (SE2)

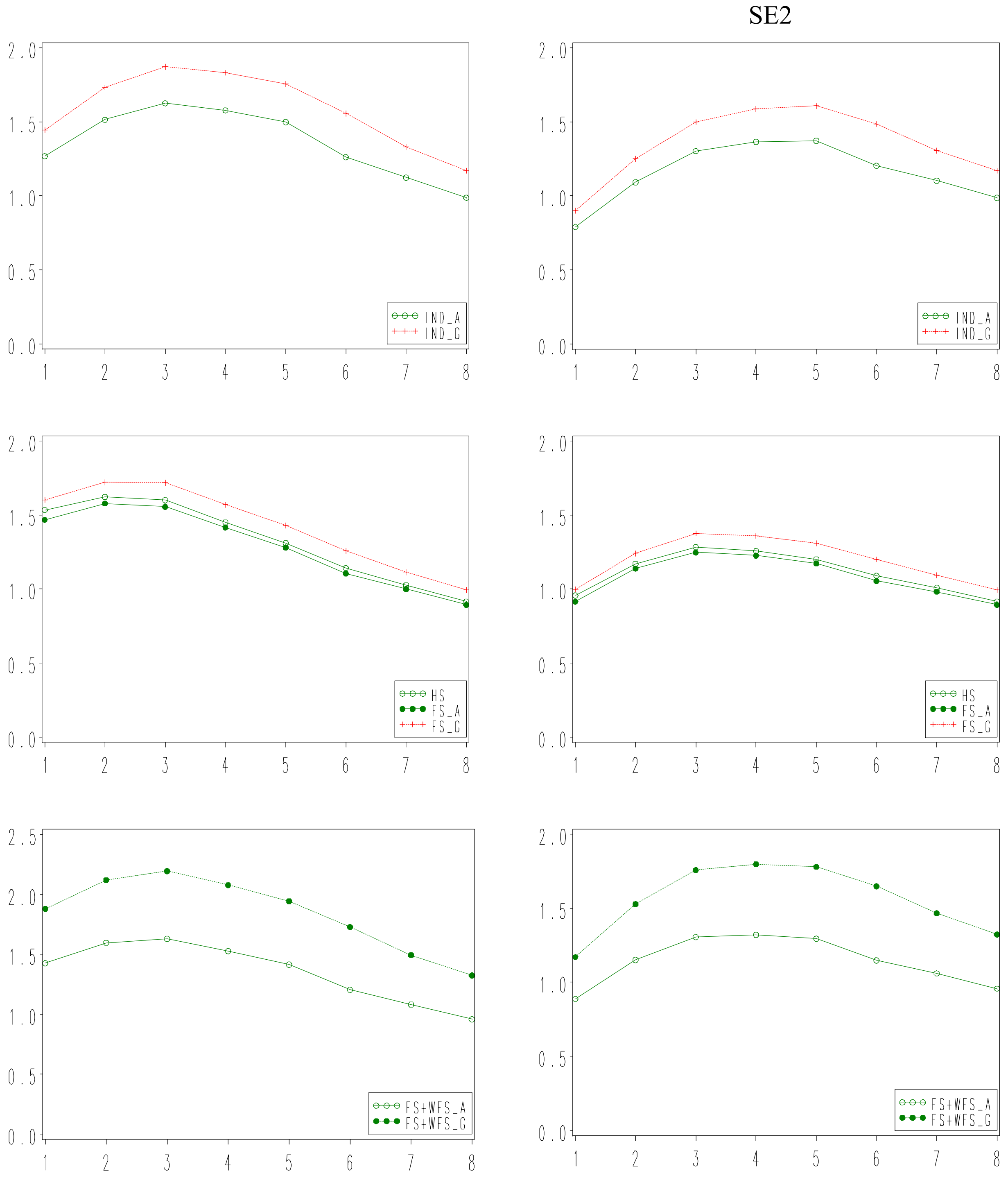

SE2

Figure 4-9. Selection efficiency of early height on 8-yr volume for Piedmont population, based on gain per year (SE1) or present value (SE2)

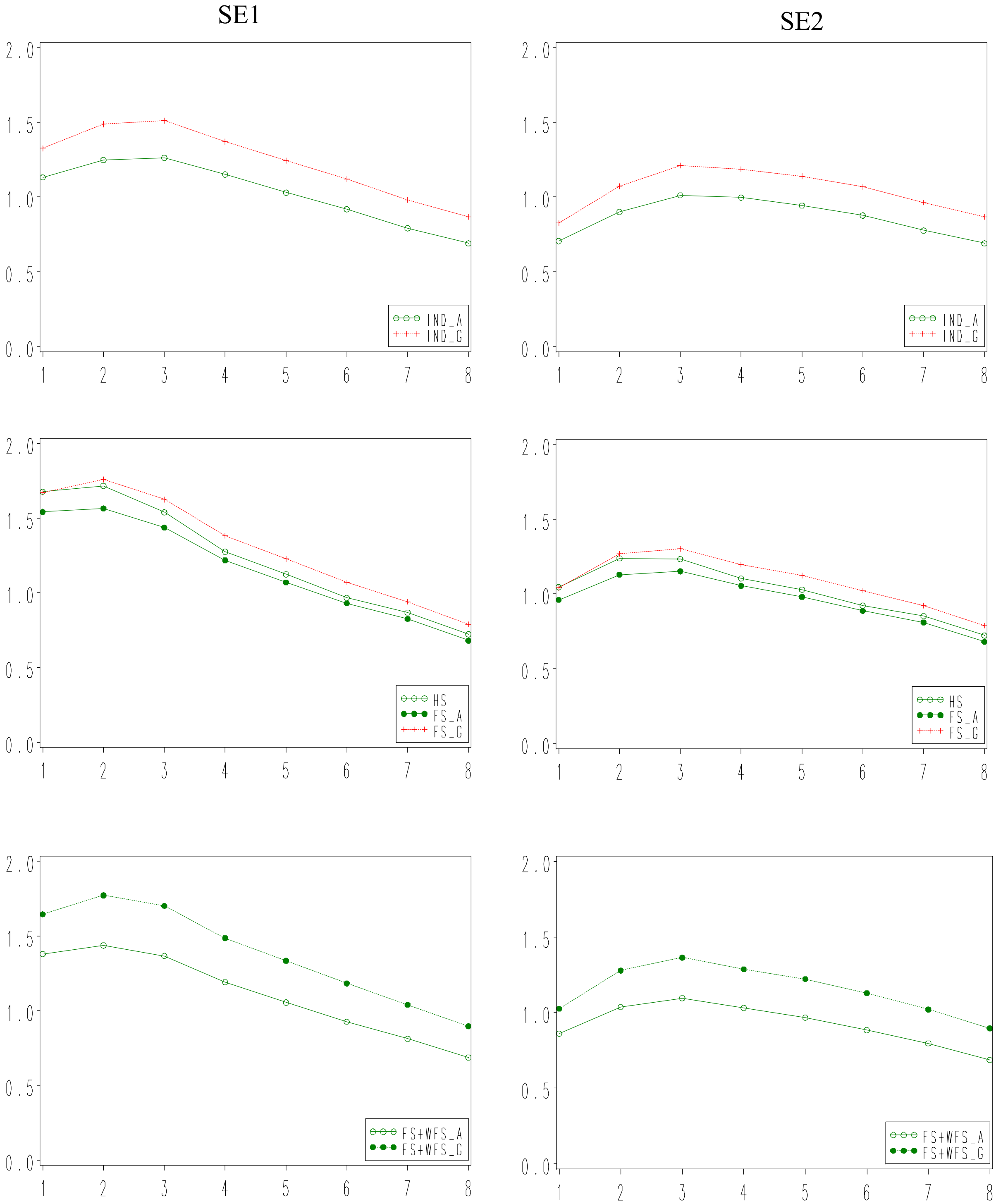


Figure 4-10. Selection efficiency of early volume on 8-yr volume for Northern population, selection efficiencies of gain per year

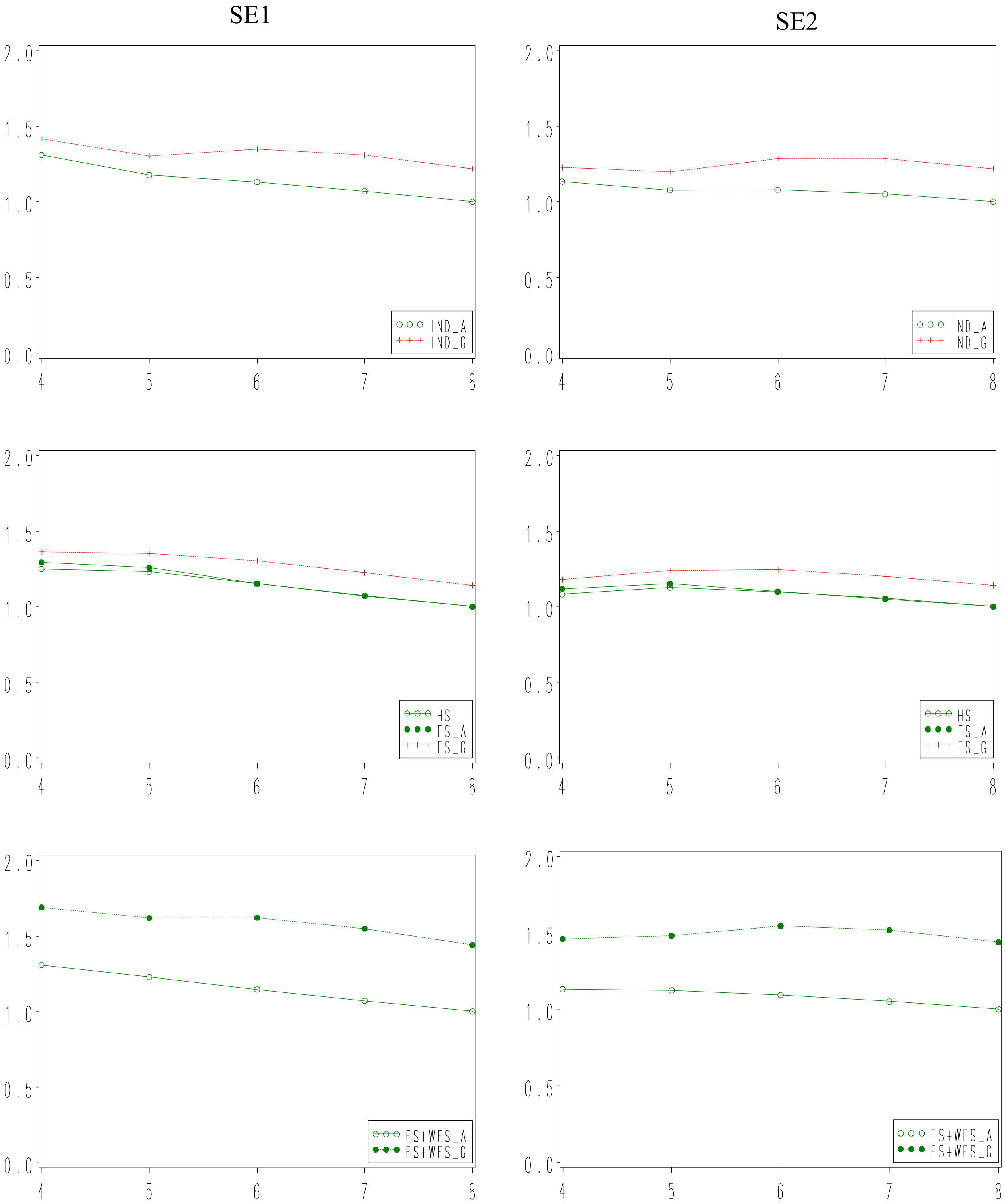


Figure 4-11. Selection efficiency of early volume on 8-yr volume for Coastal population, selection efficiencies of gain per year

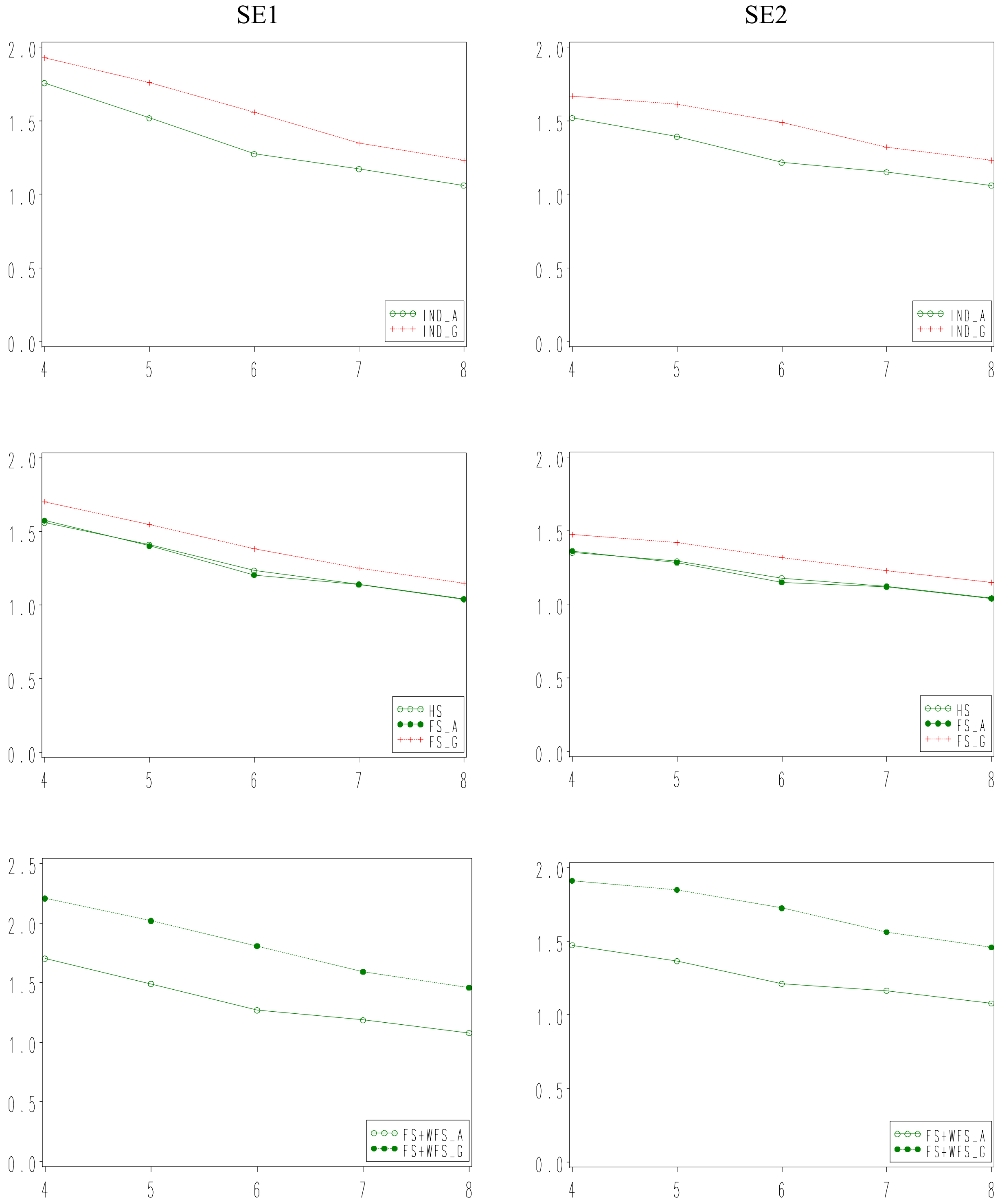
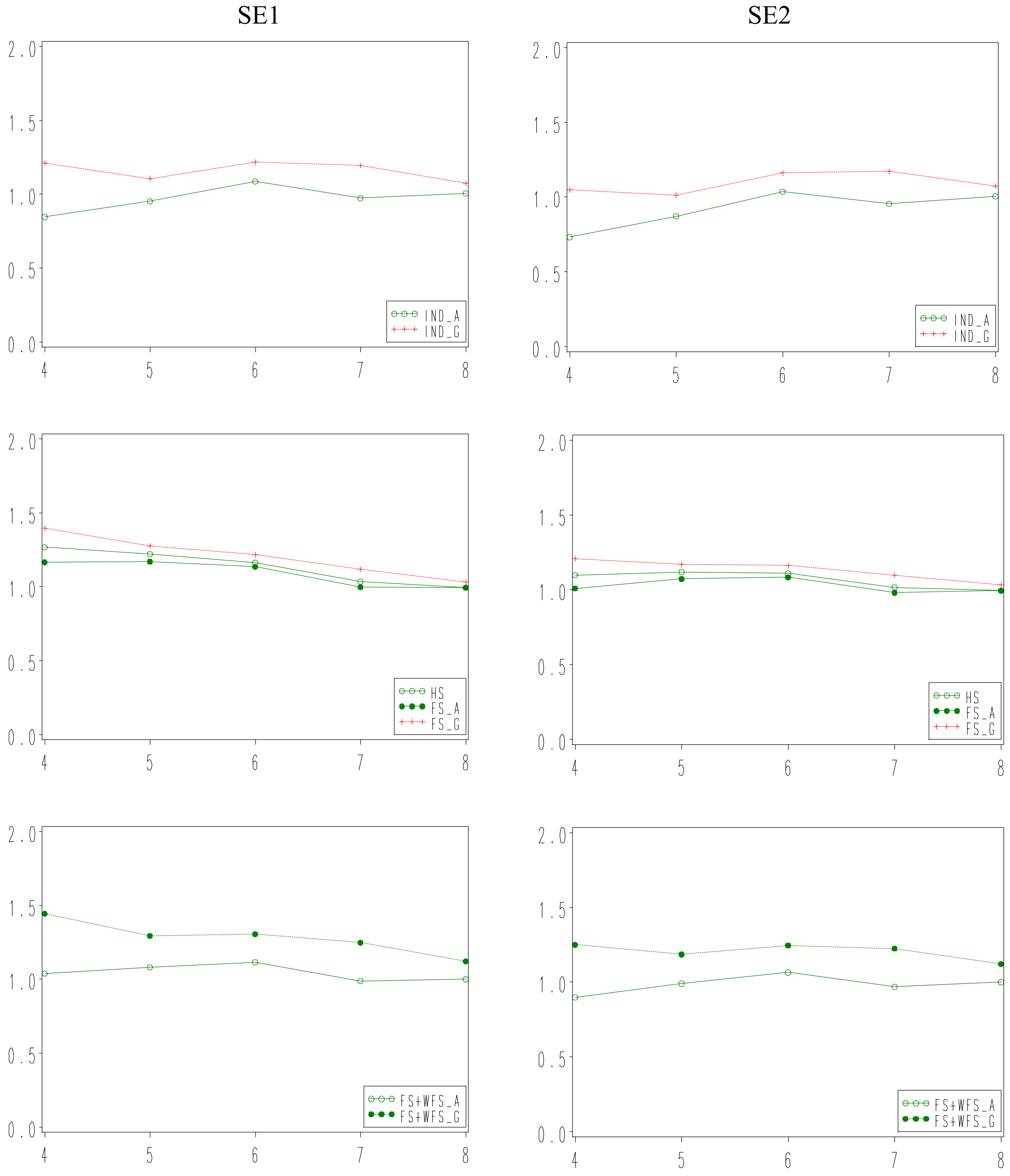


Figure 4-12. Selection efficiency of early volume on 8-yr volume for Piedmont population, selection efficiencies of gain per year


SE1
SE2
IND_A
IND_G
HS
FS_A
FS_G
FS+WFS_A
FS+WFS_G