Article

An Algorithm for Genetic Analysis of Full-Sib Datasets with Mixed-Model Software Lacking a Numerator Relationship Matrix Function, and a Comparison with Results from a Dedicated Genetic Software Package

Gary R. Hodge and Juan Jose Acosta *

Camcore, Department of Forestry & Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA; grh@ncsu.edu

* Correspondence: jjacosta@ncsu.edu

Received: 17 August 2020; Accepted: 29 October 2020; Published: 2 November 2020

Abstract: Research Highlights: An algorithm is presented that allows for the analysis of full-sib genetic datasets using generalized mixed-model software programs. The algorithm produces variance component estimates, genetic parameter estimates, and Best Linear Unbiased Prediction (BLUP) solutions for genetic values that are, for all practical purposes, identical to those produced by dedicated genetic software packages. Background and Objectives: The objective of this manuscript is to demonstrate an approach with a simulated full-sib dataset representing a typical forest tree breeding population (40 parents, 80 full-sib crosses, 4 tests, and 6000 trees) using two widely available mixed-model packages. Materials and Methods: The algorithm involves artificially doubling the dataset, so that each observation is in the dataset twice, once with the original female and male parent identification, and once with the female and male parent identities switched. Five linear models were examined: two models using a dedicated genetic software program (ASREML) with the capacity to specify A or other pedigree-related functions, and three models with the doubled dataset and a parent (or sire) linear model (ASREML, SAS Proc Mixed, and R lme4). Results: The variance components, genetic parameters, and BLUPs of the parental breeding values, progeny breeding values, and full-sib family-specific combining abilities were compared. Genetic parameter estimates were essentially the same across all the analyses (e.g., the heritability ranged from $h^2 = 0.220$ to 0.223 , and the proportion of dominance variance ranged from $d^2 = 0.057$ to 0.058). The correlations between the BLUPs from the baseline analysis (ASREML with an individual tree model) and the doubled-dataset/parent models using SAS Proc Mixed or R lme4 were never lower than $R = 0.99997$. Conclusions: The algorithm can be useful for analysts who need to analyze full-sib genetic datasets and who are familiar with general-purpose statistical packages, but less familiar with or lacking access to other software.

Keywords: diallel; variance components; breeding values; REML; BLUP; pedigree; heritability

1. Introduction

The use of linear mixed models is widespread in the analysis of genetic data, in particular, the progeny test data of animals and plants, including forest trees [1–4]. The use of the term “mixed models” refers to linear models with a combination of fixed and random effects as explanatory factors that account in varying degrees for the variation observed in a trait of interest to the breeder, e.g., milk yield in dairy cows or volume growth in forest trees. Breeders are primarily interested in predicting how specific genotypes will perform in future environments, under some type of commercial production

system for an animal product or plant crop, and therefore typically treat the underlying genetic factors affecting the trait of interest as random effects to be predicted. By contrast, breeders are typically less interested in the specific effects of the test location, or experimental replication within the test location, and they often treat these as fixed effects to be estimated [5]. Thus, the use of mixed models is well-suited for genetic data analysis, using fixed-effect estimates to adjust phenotypic data during the prediction of the random genetic effects. In particular, it is common to use software programs that combine the Restricted Maximum Likelihood (REML) estimation of variance components associated with the random effects in the model, along with use of those variance component estimates, to both estimate the fixed effects (with Best Linear Unbiased Estimates, BLUEs) and predict the random effects (Best Linear Unbiased Predictions, BLUPs).

Almost all genetic improvement programs make extensive use of designed experiments known as progeny tests. Progeny tests contain offspring arising from known or controlled matings among a set of parents, and these offspring or progeny are measured for the commercial traits of interest. In the analysis of progeny test data, a type of genetic effect of particular interest to breeders is the “additive” genetic effect. In a given population, the additive genetic effects will have a particular “additive genetic variance” associated with the distribution of those effects. This additive genetic variance (σ^2_A) represents the amount of genetic variance that can be passed from parent to progeny; thus, it is of great interest to the breeder. Alternatively, one can think of additive genetic variance as what can be inherited by the progeny from its parents; thus, the proportion of additive genetic variance relative to the total phenotypic variance is defined as the heritability (h^2) of a particular trait. The amount of genetic variance and the heritability have a huge impact on the amount of genetic gain that a breeder is likely to achieve, and on the difficulty or ease of achieving that gain [6,7].

Normally, different causal random effects in the linear model are assumed to be independently distributed, each with its own unique variance. An unusual feature of linear models for progeny test data arises from the fact that most breeding programs are focused on organisms that are diploid, which means that individuals have two copies of each gene, one inherited from the mother and one from the father. If the progeny test is a full-sib test (i.e., for all progeny, both parents are known), then an appropriately constructed linear model must somehow account for the fact that female parents and male parents are both sampling the same distribution of additive genetic effects (assuming no maternal effects or mitochondrial effects). In addition, many plant species are monoecious; that is, individual plant or tree genotypes can produce both female and male reproductive structures. Thus, in a given progeny dataset, the same parental genotype could be the female parent of some progeny and the male parent of other progeny.

A solution to deal with this data structure involves the use of a matrix $A\sigma^2_A$, which is the variance–covariance structure of the additive genetic effects for all parental and progeny genotypes in the dataset. The matrix A is known as the numerator relationship matrix, and the elements are estimates of the proportion of alleles IBD (identical by descent) in pairs of genotypes in the study [8,9]. Dedicated genetic analysis software programs typically make use of a pedigree matrix (containing a list of all progeny genotypes and their female and male parents) to form the A matrix used in the variance component estimation and BLUP of the random effects [8]. Some examples of such programs include ASREML [10], Echidna [11], MTDF-REML [12], SeleGen [13], WOMBAT [14], and others.

However, mixed-model packages or procedures are also available in a number of general-purpose statistical software packages, both commercial packages, such as SAS Proc Mixed [15], and open-source programs, such as the R lme4 package [16,17]. Most of these packages do not have a pedigree matrix or numerator relationship option, and do not have the capacity to equate female and male random terms in the linear model to represent the same additive genetic effect. The objective of this manuscript is to describe an algorithm for the genetic analysis of a full-sib progeny test dataset using general-purpose linear mixed-model software to produce variance component estimates, genetic parameter estimates, and random-effect BLUP solutions equivalent to those that are obtained using dedicated genetic data analysis software with the capacity to model the numerator relationship matrix. Such an algorithm

might be very useful for analysts familiar with general-purpose statistical packages, but less familiar with, or lacking access to, other software. The algorithm will be demonstrated with SAS Proc Mixed and R lme4

2. Materials and Methods

2.1. Overview of Analytical Approaches and Description of the Algorithm

All linear mixed models can be expressed with the following general equation:

$$y = Xb + Zu + e$$

where

- y is an $n \times 1$ vector of observations, where n is the number of records;
- X is an $n \times p$ design matrix for fixed effects, where p is the number of levels of fixed effects;
- b is the $p \times 1$ vector of fixed effects;
- Z is an $n \times q$ design matrix for random effects, where q is the number of levels of random effects;
- u is the $q \times 1$ vector of random tree effects;
- e is an $n \times 1$ vector of random residuals.

All models in this study included fixed effects for *test* (different field locations where progeny tests were established) and *replication(test)* (different replications within a test), and also included random terms for accounting for all or some proportion of the variance associated with different kinds of genetic effects: *additive*, *additive x environment*, *dominance*, and *dominance x environment*.

2.1.1. Tree Model, ASREML = ASR_ind

For this manuscript, a full-sib progeny dataset was analyzed using ASREML, with a linear model using a numerator relationship matrix. Animal breeders commonly refer to this type of analysis as the animal model, given that the numerator relationship matrix models the variance and covariances of the additive effects of individual animal genotypes in the dataset. In this manuscript, the *animal model* (or tree model, if one is a forest tree breeder) is denoted as *ASR_ind*. This was considered the baseline analysis, and four other approaches were compared to this analysis, looking at variance components associated with random effects, genetic parameters (e.g., heritability = h^2) calculated from those variance components, and BLUP solutions for random effects. Specifically, we focus on breeding value predictions (BV, the additive effect associated with a given genotype) and predictions of full-sib family Specific Combining Ability (SCA, a measure of the average dominance effect observed in the progeny of a given full-sib cross). Four other linear mixed models were compared to *ASR_ind*.

2.1.2. Parent Model, ASREML = ASR_parent

This model does not utilize a numerator relationship matrix; rather, it includes terms for the female and male parents to account for the additive effects. This model is commonly called a *sire model* by animal breeders [8] and will be referred to in this manuscript as a *parent model*, and denoted *ASR_parent*. The parent model accounts for the pedigree (parentage) of the progeny, but does not define the pedigree of the parents since they are assumed to be an unrelated base population, with their A matrix equal to an identity matrix. The *ASR_parent* model makes use of special options within ASREML to overlay the female and male design matrices, thus equating the female and male random effects. In other words, the program recognizes that those two different terms draw from the same additive genetic effect distribution, and uses them together to estimate a single variance component associated with the parent. It also recognizes that the female k and male k have the same additive effect, the one associated with parental genotype k . This model directly predicts the family General Combining Ability (GCA) [4], among the other random effect solutions. The GCA of a half-sib family k

is defined as $\frac{1}{2}$ the BV of parent k , and represents the average additive effect inherited by all progeny of parent k .

2.1.3. Progeny Breeding Value Predictions for Parent Models

The *ASR_parent* model (and the three subsequent parent models, described below) was also used to predict the BV of all progeny in the dataset as a function of the mid-parent breeding value, plus an adjustment for the Mendelian sampling of the additive genetic alleles affecting the trait of interest:

$$\hat{a}_{ijk} = \frac{1}{2} (\hat{a}_j + \hat{a}_k) + \hat{w}_{ijk}$$

where

$\hat{a}_{ijk}$ = the BV of progeny of parent $j \times$ parent k ;

$\hat{a}_j$ = the BV of parent j ;

$\hat{a}_k$ = the BV of parent k ;

$\hat{w}_{ijk}$ = the within-family Mendelian additive effect of the individual ijk .

For each tree, $\hat{w}$ was calculated as $\hat{w}_{ijk} = h^2_w (e_{ijk})$, where e_{ijk} = the residual from the linear mixed model associated with progeny from parent $j \times$ parent k , and h^2_w = the within-family heritability = $\frac{1}{2} \sigma^2_A / \sigma^2_e$, i.e., the ratio of the additive genetic variance within a full-sib family to the residual variance. This approach is effectively equivalent to the Reduced Animal Model developed by Quaas and Pollack [18], where the predicted BVs for progeny are obtained by back-solving from the parental BVs [8].

2.1.4. Algorithm for Use with General-Purpose Mixed Linear Model Software

In addition to the *ASR_ind* and *ASR_parent* models, three additional linear mixed-model scenarios were examined, all using the same linear models with exactly the same random-effect terms, but run with three different software programs (ASREML, R, and SAS). The three models are parental models, as described above, but without the use of any options or modules to force the program to equate male and female random effects. Normally, such an analysis would estimate separate and different variance components for female and male random effects, and predict different GCA values for the same parent k when used as female k or male k . The proposed algorithm for performing a proper analysis of these data does not involve modifying the mixed-model analysis programmed in the software; rather, it involves manipulating the dataset. Specifically, for any given full-sib dataset with n progeny observations, the dataset is doubled to include $2n$ total observations. Every observation is included twice, once with the true female and male parent identifiers, and again with a unique tree identifier and the female and male identifiers swapped. For example, consider two observations in a .csv data format:

```
TreeID, female, male, group, test, rep, stvolume
1241, 4, 6, B, 1, 1, 128.43
21241, 6, 4, B, 1, 1, 128.43
```

Tree 1241 is a progeny of female 4 $\times$ male 6 in test 1, rep 1 with a standardized volume = 128.43. This observation is also included in the dataset with a different TreeID (21241), as a progeny of female 6 $\times$ male 4. All the other identification variables are unchanged, as is the phenotypic measurement. The dataset is then analyzed normally with female and male random effects in the model. The program does not recognize that these effects are the same, and it will estimate separate variances for the female and male effects, but the final variance estimates produced by the software will be the same ($\sigma^2_{\text{female}} = \sigma^2_{\text{male}}$), and the BLUPs for the GCA for parent k as a female will be equal to the BLUP for GCA for parent k as a male. These results arise because in the doubled dataset, every progeny of parent k , whether truly derived with k as the female parent or k as the male parent, is included in

the dataset as progeny with k listed as the female parent (and also in another observation where k is listed as the male parent). Whether this doubling of the dataset has any other effect on the variance component estimates, genetic parameter estimates, or other BLUP solutions will be demonstrated using an example analysis.

Using this doubled dataset, parent linear mixed models were run with three software packages:

ASREML, and the model will be denoted *ASR_parDouble*;

R lme4, and the model will be denoted *R_parDouble*;

SAS Proc Mixed, and the model will be denoted *SAS_parDouble*.

2.2. Summary of Different Linear Mixed Models

The five different linear mixed models are summarized in Table 1, along with the genetic interpretations of the model variance components. Note that all the models included additive $\times$ environment interaction terms such as *female* $\times$ *environment* and *male* $\times$ *environment* in order to facilitate comparisons among the results. Additionally, note that *ASR_ind* uses the numerator relationship matrix **A**, so the variance associated with that component is the full additive genetic variance. This also means that the residual variance from that model will be lower than the residual variance in all the other models by an amount equal to $\frac{1}{2}\sigma^2_A$. For all the linear models, the following genetic parameters were then estimated as functions of the variance components:

Phenotypic variance	$= \sigma^2_{\text{phen}}$	= the sum of all the variance components in the model;
Heritability	$= h^2$	$= (\sigma^2_A / \sigma^2_{\text{phen}})$;
Proportion dominance	$= d^2$	$= (\sigma^2_D / \sigma^2_{\text{phen}})$;
Type B genetic correlation	$= r_{Bg}$	$= (\sigma^2_A / \sigma^2_A + \sigma^2_{AE})$;
Type B dominance correlation	$= r_{Bd}$	$= (\sigma^2_D / \sigma^2_D + \sigma^2_{DE})$;
Within-family heritability	$= h^2_w$	$= (1/2\sigma^2_A / \sigma^2_{\text{resid}})$.

where σ^2_A = the additive genetic variance, σ^2_{AE} = the additive $\times$ environment variance, σ^2_D = the dominance genetic variance, and σ^2_{DE} = the dominance $\times$ environment variance. The Type B genetic correlations [19] both provide a measure of the genotype $\times$ environment interaction variance, and range from $r_B = 0$, indicating no consistent genetic effect across environments, to $r_B = 1$, indicating a perfectly consistent genetic effect across all environments (or in other words, the absence of any genotype $\times$ environment variance).

Table 1. Summary of random effects included in the five different linear mixed models.

Analysis	Additive Effects			Add. $\times$ Environment		Dominance	Dom. $\times$ Env.	Residual
	TreeID	Female	Male	fe	me	fm	fme	
<i>ASR_ind</i>	x			x	x	x	x	x
<i>ASR_parent</i>		x	x	x	x	x	x	x
<i>ASR_parDouble</i>		x	x	x	x	x	x	x
<i>SAS_parDouble</i>		x	x	x	x	x	x	x
<i>R_parDouble</i>		x	x	x	x	x	x	x

TreeID $\sim N(0, \sigma^2_A)$, female $\sim N(0, \sigma^2_f)$; $\sigma^2_f = \frac{1}{4}\sigma^2_A$; male $\sim N(0, \sigma^2_m)$, $\sigma^2_m = \frac{1}{4}\sigma^2_A$; fe = female $\times$ environment, fe $\sim N(0, \sigma^2_{fe})$, $\sigma^2_{fe} = \frac{1}{4}\sigma^2_{AE}$; me = male $\times$ environment, me $\sim N(0, \sigma^2_{me})$, $\sigma^2_{me} = \frac{1}{4}\sigma^2_{AE}$; fm = female $\times$ male, fm $\sim N(0, \sigma^2_{fm})$, $\sigma^2_{fm} = \frac{1}{4}\sigma^2_D$; fme = female $\times$ male $\times$ environment, fme $\sim N(0, \sigma^2_{fme})$, $\sigma^2_{fme} = \frac{1}{4}\sigma^2_{DE}$; residual $\sim N(0, \sigma^2_{\text{resid}})$; σ^2_A = additive genetic variance; σ^2_{AE} = additive $\times$ environment variance; σ^2_D = dominance genetic variance; σ^2_{DE} = dominance $\times$ environment variance.

The specific ASREML code, R code, and SAS code for the different linear mixed models are presented as supplementary materials in Archive S1. The code and the example dataset are also available online at <https://camcore.cnr.ncsu.edu/software/>.

2.2.1. Precision of Variance Component Estimates

For all the ASREML models, the standard errors of the variance component estimates were produced directly by the program. The genetic parameters were calculated using the !VPREDICT option, and the standard errors calculated by the program. For the *SAS_parDouble* model, the ASYCOV option of Proc Mixed produces an asymptotic variance–covariance matrix of the variance component estimates. Elements of this matrix were then used to calculate the standard errors of the genetic parameters. The standard errors of h^2 were calculated using the standard error of the additive variance estimate and treating the phenotypic variance estimate as a constant according to Dickerson's approximation [20]. The standard errors of d^2 were estimated in a similar manner. The standard errors of the Type B correlations (r_{Bg} and r_{Bd}) were calculated as a first-order approximation of a Taylor-expansion series [21].

The R package lme4 does not provide standard error estimates for the variance components, but does provide confidence interval estimates around the variance component estimates. This is a philosophical decision by the program developer, reflecting the view that standard errors should be restricted for use with normal or nearly normal distributions, and thus are not appropriate for variance component estimates that are asymmetrically distributed [22]. Using the *confint* function of lme4, a 68.3% confidence interval was calculated for the variance component estimates of the *R_parDouble* model. This confidence interval covers the same range as a ± 1.0 standard deviation range would for a normally distributed variable.

2.2.2. Precision of BLUP Estimates

The standard errors of prediction for the BLUPs for the parental BVs ($SE(a - \hat{a})$) and full-sib SCA effects ($SE(s - \hat{s})$) were directly calculated from the software for all the linear models. The standard errors of prediction for the BLUPs for the progeny BVs were directly calculated only for the *ASR_ind* model. For the other models, it is theoretically possible to estimate $SE(a - \hat{a})$ for progeny (e.g., as functions of the $SE(a - \hat{a})$ of the parents, covariances among the $\hat{a}$ of the parents, the within-family heritability, and the within-family additive genetic variance), but these are not produced directly by the software.

2.2.3. Correlations among BLUPs from Different Models

To compare BLUPs from the different linear models, we examined the correlations among the genetic value predictions $\text{Corr}(\hat{a}_k, \hat{a}_{ASR_ind})$ and developed linear regression models relating the predicted genetic values from the four parent models ($\hat{a}_k$) to those from *ASR_ind* ($\hat{a}_{ASR_ind}$).

2.2.4. Description of the Dataset

A simulated dataset representative of a typical forest tree-breeding scenario was used to compare the different linear mixed models. Specifically, the initial genetic parameters were assumed based on results typical of the 8-year stem volume for tropical pines [23]: $h^2 = 0.20$, $d^2 = 0.10$, $r_{Bg} = 0.75$, $r_{Bd} = 0.75$, and total phenotypic variance = 2500. Using these genetic values, a population of 40 parents was generated, and these 40 parents were crossed in an incomplete diallel to form 80 full-sib families. The different parents were included in different numbers of crosses: of the 40 parents, 4 parents were included in only 2 crosses, 8 parents were used in 3 crosses, 16 parents in 4 crosses, 8 parents in 5 crosses, and 4 parents were used in 6 crosses. The 80 full-sib families were divided into four sets (A, B, C, and D) and established in four field tests. Each field test was assumed as a randomized complete block with 25 replications, with single-tree plots (one tree/family/replication). All the full-sib families were included in three of the four tests, with each test containing three of the sets (i.e., ABC, ABD, ACD, and BCD) in Tests 1 to 4, respectively. For the purposes of the simulation, 100% survival was assumed, so there were 6000 progeny from 40 parents and 80 full-sib families in the total dataset.

All the simulated data were generated using a modification of the algorithm described by Lstiburek et al. [24].

3. Results

3.1. Variance Component Estimates

Throughout the presentation of the results and the following discussion, for convenience, we will use the estimates from the *ASR_ind* model as the “best” estimates, and compare the other models to these. Estimates of the variance components from the five linear models are presented in Table 2. The two ASREML models using the original dataset (*ASR_ind* and *ASR_parent*) gave identical variance component estimates for all the terms (up to the fifth significant digit). All of the models using the doubled dataset produced very similar estimates for all model terms. For the additive genetic variance, *ASR_ind* produced an estimate of $\hat{\sigma}_A^2 = 557.5$, and the range of the three doubled-dataset models was from $\hat{\sigma}_A^2 = 555.5$ to 557.4. For the female $\times$ male variance (σ_{fm}^2), which can be interpreted as the variance of the full-sib SCA effects, *ASR_ind* produced an estimate of $\hat{\sigma}_{fm}^2 = 36.5$, and the range of the three doubled-dataset models was from $\hat{\sigma}_{fm}^2 = 35.6$ to 35.9. The female $\times$ environment and female $\times$ male $\times$ environment interaction variances (σ_{fe}^2 and σ_{fme}^2) were also very similar across the models. *ASR_ind* produced an estimate of $\hat{\sigma}_{fe}^2 = 47.3$, and the range of the three doubled-dataset models was $\hat{\sigma}_{fe}^2 = 44.7$ to 45.9. *ASR_ind* produced an estimate of $\hat{\sigma}_{fme}^2 = 25.0$, and the range of the three doubled-dataset models was $\hat{\sigma}_{fme}^2 = 25.9$ to 27.1. Finally, *ASR_ind* gave an estimate of the phenotypic variance of $\hat{\sigma}_{phen}^2 = 2534.5$, and the range of the three doubled-dataset models was $\hat{\sigma}_{phen}^2 = 2495.5$ to 2513.4.

Table 2. Estimated variance components ¹ and standard errors ² from five linear mixed models ³ for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test.

Analysis	Estimated Variance Components													
	σ^2_A	<i>se</i>	σ^2_f	<i>se</i>	σ^2_{fe}	<i>se</i>	σ^2_{fm}	<i>se</i>	σ^2_{fme}	<i>se</i>	σ^2_{resid}	<i>se</i>	σ^2_{phen}	<i>se</i>
<i>ASR_ind</i>	557.5	167.2	139.4	41.8	47.3	13.8	36.5	18.3	25.0	19.2	1820.9	92.4	2534.5	92.4
<i>ASR_parent</i>	557.5	167.2	139.4	41.9	47.3	13.8	36.5	18.3	25.0	19.2	2099.5	39.5	2534.4	92.4
<i>ASR_parDouble</i>	555.8	163.7	139.0	40.9	45.9	13.5	35.9	12.8	25.9	13.5	2081.9	27.5	2513.4	87.4
<i>SAS_parDouble</i>	555.5	163.7	138.9	40.9	45.9	13.7	35.9	12.9	25.9	13.7	2081.9	27.5	2513.2	65.3
<i>R_parDouble</i>	557.4	−149.2, +176.4	139.4	−37.3, +44.1	44.7	−13.0, +13.8	35.6	−11.9, +14.2	27.1	−12.7, +15.1	2064.7	−27.1, +27.3	2495.5	−12.7, +15.1

¹ The *ASR_ind* model included a term for an individual tree, so the software directly estimated additive variance (σ^2_A), and female parent variance ($\sigma^2_f = \frac{1}{4} \sigma^2_A$) was derived. All other models included terms for female and male parents, so the software directly estimated female parent variance ($\sigma^2_f = \frac{1}{4} \sigma^2_A$), and additive variance (σ^2_A) was derived. Other model terms were female x environment (σ^2_{fe}), female x male (σ^2_{fm}), female x male x environment (σ^2_{fme}), residual (σ^2_{resid}), and total phenotypic variance (σ^2_{phen}). ² The R package lme4 does not provide standard error estimates for variance component estimates; rather, it provides confidence intervals around the estimate. Values presented are for a 68.3% confidence interval, equivalent to ± 1.0 standard deviation for a normally distributed variable. ³ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R lme4), linear model terms (individual tree or parent model), and/or the use of a normal or doubled dataset.

3.2. Precision of Variance Component Estimates

As expected, the estimated standard errors of the variance components from *ASR_ind* and *ASR_parent* were identical (with the exception of $se(\hat{\sigma}_{resid}^2)$, which makes sense since σ_{resid}^2 is different in the two models). Comparing the standard errors of the variance component estimates from the doubled-dataset models to those from the *ASR_ind* and *ASR_par* models, there are lower standard error estimates for some of the random effects (Table 2). The standard error estimates from *ASR_ind* and the doubled-dataset models (*ASR_parDouble* and *SAS_parDouble*) were similar for the additive variance, with $se(\hat{\sigma}_A^2) = 167.2$ compared to $se(\hat{\sigma}_A^2) = 163.7$ and 163.7 , respectively. This was also observed for the *female* $\times$ *environment* variance, with $se(\hat{\sigma}_{fe}^2)$ estimates of 13.8 compared to 13.5 and 13.7 from *ASR_ind*, *ASR_parDouble*, and *SAS_parDouble*, respectively. For the additive variance, *R_parDouble* estimated $\hat{\sigma}_A^2 = 557.4$, with a 68.3% confidence interval from $\hat{\sigma}_A^2 - 149.2$ to $\hat{\sigma}_A^2 + 176.4$. This is comparable to the range of estimated $\hat{\sigma}_A^2 = 557.5 \pm 167.2$ from the *ASR_ind* model (Table 2).

There was, however, a notable difference in the standard error estimates for the dominance-related effects, the *female* $\times$ *male* and *female* $\times$ *male* $\times$ *environment* variance, and to a lesser extent, the residual variance (Table 2). For the *female* $\times$ *male* effect, the $se(\hat{\sigma}_{fm}^2)$ from *ASR_ind* was $se(\hat{\sigma}_{fm}^2) = 18.3$ versus $se(\hat{\sigma}_{fm}^2) = 12.8$ and 12.9 from *ASR_parDouble* and *SAS_parDouble*, respectively. Similarly, for the *female* $\times$ *male* $\times$ *environment* effect, *ASR_ind* gave $se(\hat{\sigma}_{fme}^2) = 19.2$ versus $se(\hat{\sigma}_{fme}^2) = 13.5$ and 13.7 from *ASR_parDouble* and *SAS_parDouble*, respectively. Finally, for the *residual* variance, the $se(\hat{\sigma}_{resid}^2)$ from *ASR_par* was $se(\hat{\sigma}_{resid}^2) = 39.5$, while the estimate from the doubled-dataset analysis with *ASR_parDouble* was $se(\hat{\sigma}_{resid}^2) = 27.5$. For these variance component estimates from the *R_parDouble* model, the confidence intervals were also more narrow than would be expected compared to the results from the *ASR_ind* model (Table 2).

This apparent difference in the precision of these variance component estimates from the doubled-dataset analyses is due to fact that the software “counts” twice as many degrees of freedom to estimate the *female* $\times$ *male*, *female* $\times$ *male* $\times$ *environment*, and *residual* variances as are actually present in the original dataset. Specifically, for this example, there are 80 full-sib (*female* $\times$ *male*) crosses in the dataset, and 240 full-sib $\times$ environment combinations (each full-sib family tested on three sites). However, because of the doubling of the dataset, each unique *female* $\times$ *male* combination is also a unique *male* $\times$ *female* when the parents are reversed. Thus, the software “counts” 160 full-sib (*female* $\times$ *male*) crosses in the dataset, and 480 full-sib $\times$ environment combinations, and this leads to lower standard errors (or a smaller confidence interval, in the case of *R_parDouble* using *R lme4*) for these variance components.

3.3. Genetic Parameter Estimates

Although there was some slight variation among the variance component estimates from the different models, there was no important variation in the genetic parameter estimates from the different models (Table 3). *ASR_ind* and *ASR_parent* produced identical genetic parameter estimates. For most of the genetic parameters, the three doubled-dataset models produced estimates that differed from the *ASR_ind* estimates only at the third significant figure. For example, for heritability, *ASR_ind* gave an estimate of $\hat{h}^2 = 0.220$, and the range of the three doubled-dataset models was $\hat{h}^2 = 0.221$ to 0.223 . For the proportion of dominance, *ASR_ind* gave an estimate of $\hat{d}^2 = 0.058$, and all three doubled-dataset models produced an estimate of $\hat{h}^2 = 0.057$. For the Type B genetic correlation, *ASR_ind* gave an estimate of $\hat{r}_{Bg} = 0.747$, and the range of the three doubled-dataset models was $\hat{r}_{Bg} = 0.752$ to 0.757 . For the Type B dominance correlation, *ASR_ind* gave an estimate of $\hat{r}_{Bd} = 0.594$, and the range of the three doubled-dataset models was $\hat{r}_{Bd} = 0.568$ to 0.581 . Among the genetic parameter estimates, this was the largest deviation observed from the *ASR_ind* parameter estimate. Finally, for the within-family heritability (h_w^2), *ASR_ind* gave an estimate of $\hat{h}_w^2 = 0.133$, and the range of the three doubled-dataset models was $\hat{h}_w^2 = 0.133$ to 0.135 . This is an important parameter, because it is used to calculate the within-family additive genetic deviation from the mid-parent breeding value, the Mendelian genetic sampling term.

Table 3. Estimated genetic parameters ¹ and standard errors ² from five linear mixed models ³ for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test.

Analysis	Genetic Parameter Estimates									
	h^2	<i>se</i>	d^2	<i>se</i>	r_{Bg}	<i>se</i>	r_{Bd}	<i>se</i>	h^2_w	<i>se</i>
<i>ASR_ind</i>	0.220	0.059	0.058	0.029	0.747	0.082	0.594	0.254	0.133	0.040
<i>ASR_parent</i>	0.220	0.059	0.058	0.029	0.747	0.082	0.594	0.254	0.133	0.040
<i>ASR_parDouble</i>	0.221	0.058	0.057	0.020	0.752	0.081	0.581	0.177	0.134	0.039
<i>SAS_parDouble</i>	0.221	0.065	0.057	0.020	0.752	0.082	0.581	0.179	0.133	0.039
<i>R_parDouble</i>	0.223	−0.060, +0.071	0.057	−0.019, +0.023	0.757		0.568		0.135	−0.036, +0.043

¹ Genetic parameter estimates are h^2 = narrow-sense heritability, d^2 = proportion of dominance, r_{Bg} = Type B additive genetic correlation, and r_{Bd} = Type B dominance genetic correlation.

² The R package *lme4* does not provide standard error estimates for variance component estimates; rather, it provides confidence intervals around the estimate, so standard errors of the genetic parameters were not estimated. Confidence intervals for r_{Bg} = Type B additive genetic correlation and r_{Bd} = Type B dominance genetic correlations were not calculated. ³ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R *lme4*), linear model terms (individual tree or parent model), and/or the use of a normal or doubled dataset.

3.4. BLUPs of Genetic Effects

Tables 3–5 present a subset of the predicted breeding values for the parents, the SCA values for the full-sib families, and the breeding values for the progeny. The selected subsets cover the range of the genetic value predictions, sampling approximately every $\frac{1}{2}$ standard deviation in the population from minimum to maximum (based on the z-scores for the *ASR_ind* predictions).

For the parental breeding value predictions for the parents ($\hat{a}$), the five different linear models produced nearly identical predictions for particular genotypes (Table 4). For any genotype with a breeding value prediction more than $\frac{1}{2}$ a standard deviation away from the population mean of zero, the predictions from the different linear models differed only at the third significant figure. For example, for Parent 7, which had a z-score around 2.2 among the parental breeding value predictions, the predictions from the five linear models were $\hat{a} = 44.53, 44.54, 44.62, 44.76$, and 44.60 .

Table 4. Parent breeding value predictions ($\hat{a}$) from five linear mixed models ¹ for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test. $SD(\hat{a})$ = the observed standard deviation of the population of 40 breeding value predictions. $SE(a - \hat{a})$ = the average standard error of prediction ². Selected parents span the distribution of 40 parents; standardized z-values are listed ³.

ParentID	ASR_ind	ASR_parent	ASR_parDouble	SAS_parDouble	R_parDouble	z
20	47.50	47.50	47.64	47.64	47.78	2.30
7	44.53	44.54	44.62	44.60	44.76	2.16
22	32.55	32.54	32.68	32.66	32.82	1.58
27	19.33	19.33	19.38	19.38	19.38	0.94
14	13.81	13.82	13.84	13.84	13.88	0.67
10	−0.22	−0.22	−0.19	−0.19	−0.12	−0.01
16	−10.55	−10.55	−10.58	−10.58	−10.58	−0.51
34	−20.69	−20.70	−20.78	−20.78	−20.83	−1.00
36	−47.41	−47.42	−47.56	−47.54	−47.67	−2.30
mean	0.00	0.00	0.00	0.00	0.00	
$SD(\hat{a})$	20.65	20.65	20.71	20.70	20.76	
$SE(a - \hat{a})$	11.87	11.87	11.71	11.71	11.71	

¹ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R lme4), linear model terms (individual tree or parent model), and/or the use of a normal or modified doubled dataset. ² Standard error of prediction for each of the 40 parents was estimated directly by the software. ³ z-values based on breeding value (BV) predictions from model *ASR_ind*.

For the full-sib SCA predictions ($\hat{s}$), the five different linear models produced nearly identical predictions for particular genotypes (Table 5). For any full-sib cross with a z-score more than 0.50, the predictions from the different linear models differed only at the third significant figure. For example, for the full-sib family of female 28 × male 21, with a z-score = −2.17 for the predicted SCA effect, the predictions from the five linear models were $\hat{s}_{28_21} = -6.92, -6.91, -6.86, -6.83$, and -6.87 (female_male = 28_21 in Table 5).

Table 5. Full-sib family Specific Combining Ability (SCA) predictions ($\hat{s}$) from five linear mixed models ¹ for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test. $SD(\hat{s})$ = the observed standard deviation of the population of 80 SCA predictions. $SE(a - \hat{a})$ = the average standard error of prediction ². Selected full-sib families have SCA values that span the distribution of 80 families; standardized z-values are listed. ³.

Female_Male	ASR_ind	ASR_parent	ASR_parDouble	SAS_parDouble	R_parDouble	z
29_28	9.26	9.26	9.19	9.19	9.13	2.90
10_7	7.26	7.26	7.19	7.19	7.15	2.28
12_9	4.52	4.52	4.48	4.48	4.45	1.42
32_29	3.26	3.26	3.23	3.23	3.23	1.02
16_15	1.61	1.61	1.60	1.60	1.59	0.50
4_3	0.06	0.06	0.07	0.06	0.10	0.02
31_30	−1.46	−1.46	−1.46	−1.46	−1.47	−0.46
37_36	−3.15	−3.15	−3.12	−3.12	−3.12	−0.99
12_5	−5.25	−5.25	−5.20	−5.20	−5.14	−1.65
28_21	−6.92	−6.91	−6.86	−6.87	−6.83	−2.17
24_23	−7.34	−7.34	−7.26	−7.26	−7.21	−2.30
mean	0.00	0.00	0.00	0.00	0.00	
SD($\hat{s}$)	3.19	3.19	3.16	3.16	3.14	
SE($s - \hat{s}$)	5.15	5.15	5.10	5.10	5.11	

¹ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R lme4), linear model terms (individual tree or parent model), and/or the use of a normal or modified doubled dataset. ² Standard error of prediction for each of the 80 SCA effects was estimated directly by the software. ³ z-values based on BV predictions from model ASR_ind.

Finally, as with the parental breeding value predictions, for the progeny breeding value predictions ($\hat{a}$), the five different linear models produced nearly identical predictions for particular genotypes (Table 6). For any progeny with a z-score more than 0.50, the predictions from the different linear models differed only at the third significant figure. For example, for Tree 1692, with a z-score = +3.00 for the predicted BV, the predictions from the five linear models were $\hat{a} = 46.70, 46.70, 46.82, 46.97,$ and 46.80. Since there were 6000 progeny, the z-score of the predicted breeding values covered a wider range ($z = -2.80$ to $+3.86$) than was observed among the 40 parents ($z = -2.30$ to $+2.30$).

Table 6. Individual tree-breeding value predictions from five linear mixed models ¹ for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test. $SD(\hat{a})$ = the observed standard deviation of the population of 6000 breeding value predictions. $SE(a - \hat{a})$ = the average standard error of prediction ². Selected trees have breeding values that span the distribution of 6000 trees; standardized z-values are listed ³.

treeID	ASR_ind	ASR_parent	ASR_parDouble	SAS_parDouble	R_parDouble	Z
1693	60.02	60.03	60.21	60.20	60.53	3.86
1692	46.70	46.70	46.82	46.80	46.97	3.00
1663	38.96	38.96	39.04	39.04	39.15	2.50
6992	31.18	31.18	31.25	31.25	31.28	2.00
7198	23.50	23.50	23.56	23.56	23.59	1.50
15715	15.76	15.76	15.86	15.85	15.95	1.00
9695	8.03	8.03	8.04	8.03	8.01	0.50
2752	0.31	0.31	0.33	0.33	0.38	0.00
3596	−7.42	−7.42	−7.44	−7.44	−7.44	−0.50

Table 6. Cont.

treeID	ASR_ind	ASR_parent	ASR_parDouble	SAS_parDouble	R_parDouble	Z
2663	−15.17	−15.17	−15.23	−15.22	−15.34	−1.00
3204	−22.88	−22.87	−22.93	−22.93	−22.99	−1.50
2557	−30.65	−30.64	−30.72	−30.71	−30.79	−2.00
3156	−38.33	−38.34	−38.45	−38.44	−38.68	−2.50
13308	−42.99	−42.99	−43.15	−43.13	−43.40	−2.80
mean	0.30	0.30	0.31	0.30	0.31	
SD($\hat{a}$)	15.45	15.45	15.50	15.50	15.56	
SE($a - \hat{a}$)	17.37					

¹ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R lme4), linear model terms (individual tree or parent model), and/or the use of a normal or modified doubled dataset. ² Standard error of prediction for each of the 6000 progeny was estimated directly by the software only with the ASR_ind model.

³ z-values based on BV predictions from model ASR_ind.

3.5. Precision of the BLUPs

Similar to the results with the BLUP solutions from the different linear models, the estimated standard errors of the predictions from the different linear models were very similar. For example, the average standard error of the predictions for the 40 parents from the ASR_ind model was $SE(a - \hat{a}) = 11.87$, compared to $SE(a - \hat{a}) = 11.71$ for the three doubled-dataset models (Table 4). For the SCA effects, the average standard error of the predictions for the 80 full-sib families from the ASR_ind model was $SE(a - \hat{a}) = 5.15$, while $SE(a - \hat{a})$ ranged from 5.09 to 5.11 for the three doubled-dataset models (Table 5). Comparing the $SE(a - \hat{a})$ for the parents and progeny, the average $SE(a - \hat{a}) = 17.37$ for the progeny (Table 6) compared to the average $SE(a - \hat{a}) = 11.87$ for the parents (Table 4), and every progeny had a larger $SE(a - \hat{a})$ than every parent (see supplemental materials, Archive S1).

As data increase in quality and quantity, BLUP predictions for a given random effect will have not only lower errors of prediction (e.g., $SE(a - \hat{a})$), but also a higher variance among predictions, approaching the estimated variance for that random effect [8,25,26]. The breeding value predictions for both parents and progeny are scaled based on the additive genetic variance. From the ASR_ind model, the estimated additive variance was $\hat{\sigma}_A^2 = 557.5$, which corresponds to $\hat{\sigma}_A = 23.6$, which estimates the spread among the “true” breeding values in the population. Since, in this dataset, the parental breeding values were based on many progeny, they were precisely predicted, and the distribution of the parental breeding value predictions approached the distribution of the true breeding values. In fact, the standard deviation of the parent breeding value predictions was $SD(\hat{a}) = 20.65$ for ASR_ind, with similar values for the other models (Table 4). For the progeny, the standard deviation of the breeding value predictions was smaller, $SD(\hat{a}) = 15.45$ for ASR_ind, with similar values for the other models (Table 6). For the progeny, one-half of the breeding value arises from Mendelian sampling within the full-sib family, and the only datum for predicting that component is the single phenotypic measurement on that tree, which in this study, was relatively imprecise (i.e., $h^2_w = 0.133$, Table 3). Thus, it makes sense that individual tree BVs are predicted with less precision than parent BVs and therefore span a smaller range. There is a larger observed range among the progeny BVs (from −42.99 to 60.02, Table 6) compared to the parental BVs (−47.41 to 47.50, Table 5); however, this is entirely due to the fact that there were only 40 parents sampled compared to 6000 progeny. With such a large number of progeny, a few extreme outliers were found (z-scores of $z = -2.80$ and $z = 3.86$, Table 6), but most of the progeny BVs were clustered closer around the mean of zero.

The predictions of the SCA effects are scaled based on $\frac{1}{4}\sigma_D^2$, which is estimated by $\hat{\sigma}_{fm}^2$ in the linear models of Table 2. For model ASR_ind, $\hat{\sigma}_{fm}^2 = 36.5$, corresponding to $\hat{\sigma}_{fm} = 6.0$, which estimates the spread among the “true” SCA effects in the population. In this simulated experiment, full-sib families were tested on three sites with 25 progeny per site, and the $SD(\hat{s})$ was around 3.2 for all five models.

3.6. Correlations among Genetic Value Predictions from Different Models

Perhaps most importantly, there was an extremely high correlation among the predicted genetic values from all five linear models, for the parental breeding value predictions, progeny breeding value predictions, and full-sib family SCA predictions (Table 7). The correlation between the genetic values predicted from the four parent linear models and the *ASR_ind* model was never below $\text{Corr}(\hat{a}_m, \hat{a}_{ASR_ind}) = 0.99997$ and was mostly 1.00000. The linear regression models relating $\hat{a}_m$ to $\hat{a}_{ASR_ind}$ (or $\hat{s}_m$ to $\hat{s}_{ASR_ind}$) had slopes very near to 1.00, and intercepts very near to zero (Table 6). It is also worth noting that the differences between $\hat{a}_m$ and $\hat{a}_{ASR_ind}$ (or $\hat{s}_m$ and $\hat{s}_{ASR_ind}$) are quite small. For example, for the breeding values of the progeny, the model *R_parDouble* had the largest differences between $\hat{a}_m$ and $\hat{a}_{ASR_ind}$, with a standard error of the differences of $SD(\hat{a}_m - \hat{a}_{ASR_ind}) = 0.1284$. The range among the progeny breeding value predictions of $SD(\hat{a}) = 15.45$ (Table 7), so in this case, the ratio of $SD(\hat{a}_m - \hat{a}_{ASR_ind})/SD(\hat{a}_m)$ is 0.0083. In other words, relative to the variation among the predicted genetic values, the difference in the predictions from the different methods is practically negligible.

Table 7. Relationships between BLUPs from *ASR_ind* and other linear mixed model BLUPs¹ for parental breeding value, full-sib family SCA, and individual tree breeding value. Statistics are R = correlation, linear regression slope, linear regression intercept, and $se(\hat{g}_k - \hat{g}_{ASR_ind})$ = standard error of difference between BLUPs, where $\hat{a}$ = breeding value prediction, $\hat{s}$ = SCA prediction. Analyses for a simulated full-sib progeny dataset with 40 parents, 80 families from an unbalanced mating design, 4 tests, and 25 trees/family/test.

BV Progeny				
Model	R	Slope	Intercept	$se(\hat{a}_m - \hat{a}_{ASR_ind})$
<i>ASR_parent</i>	1	1.000012	−0.000169	0.0037
<i>ASR_parDouble</i>	1	1.003019	0.001407	0.0524
<i>SAS_parDouble</i>	1	1.002756	0.000972	0.0478
<i>R_parDouble</i>	0.99999	1.006739	0.003667	0.1284
BV Parent				
Model	R	slope	intercept	$se(\hat{a}_m - \hat{a}_{ASR_ind})$
<i>ASR_parent</i>	1	1.000042	0.000248	0.0048
<i>ASR_parDouble</i>	1	1.002849	0.000637	0.0649
<i>SAS_parDouble</i>	1	1.002607	0.000419	0.0599
<i>R_parDouble</i>	1	1.005545	0.000420	0.1286
SCA Family				
Model	R	slope	intercept	$se(\hat{s}_m - \hat{s}_{ASR_ind})$
<i>ASR_parent</i>	1	0.999766	0.000047	0.0009
<i>ASR_parDouble</i>	0.99999	0.990685	0.000047	0.0317
<i>SAS_parDouble</i>	0.99999	0.990869	0.000019	0.0310
<i>R_parDouble</i>	0.99997	0.984316	0.000016	0.0557

¹ The analyses varied according to the software (ASREML, SAS Proc Mixed, or R lme4), linear model terms (individual tree or parent model), and/or the use of a normal or modified doubled dataset.

4. Discussion

4.1. Genetic Value Predictions

Perhaps the most important objective in the analysis of genetic datasets is the prediction of genetic values to guide selection. In progeny test analysis, accurate and precise predictions of breeding values are critically important. Breeding values represent additive genetic value, so these predictions guide selection decisions among parents, progeny, or both, in the formation of a base population to begin the next cycle of breeding and improvement. Breeding value predictions are used to make selection decisions in the formation of a production population, that is, those genotypes that will be used

to produce progeny for commercial deployment [5]. In forest trees, the most common form of the production population is a clonal seed orchard (CSO) (e.g., see [27,28]). In a CSO, a small number of selected genotypes are multiplied through some type of vegetative propagation and established in a single location. Following open pollination due to wind or insects, seed collected from the orchard is used to establish commercial plantings. In some cases, rather than relying on open pollination, specific full-sib families are selected as the deployment option [29]. The full-sib families are produced by controlled pollination, producing a large amount of seed, with seedlings as the deployment unit, or a small amount of seed that is multiplied vegetatively to produce rooted cuttings as the deployment unit. In this case, the genetic value of the full-sib family would be equal to the average of the two parental breeding values, plus the SCA effect for that family. In this case, accurate and precise predictions of the SCA effects would be important for making optimal selection decisions among all the possible full-sib families.

The doubled-dataset/parent-model algorithm described in this manuscript produced essentially identical breeding value predictions to an individual tree analysis for parental and progeny breeding value predictions, and SCA effects for full-sib families. In every case, the correlation among the predictions from the *ASR_ind* model and the doubled-dataset models was essentially unity, the linear regression coefficients were essentially one, and the linear regression intercept was essentially zero. These data indicate that the breeding values for parents, SCA effects for full-sib families, and breeding values for progeny from the four parent models are essentially identical to those from *ASR_ind*; thus, regardless of which linear model is used, there will be no differences in terms of selection decisions, and no differences in genetic gain predictions for a population of selected genotypes.

4.2. Variance Component Estimates

Some small differences in the variance component estimates produced by the different linear models and methods might be expected due to differences among the software in the convergence algorithms. ASREML uses the Average Information algorithm and sparse matrix methods [30,31] and assesses convergence by looking at the change in the REML log-likelihood [32]. SAS Proc Mixed uses the Newton–Raphson algorithm and assesses convergence by looking at the change in the REML log-likelihood [15,32]. The R lme4 package uses a hybrid approach, beginning with a number of iterations using the Expectation-Maximization algorithm [33], followed by a number of Newton–Raphson iterations, and assesses convergence by looking at the relative change in the variance parameter estimates [32]. These differences in convergence algorithm among the three software packages (and perhaps other differences such as the precision of the calculations) could account for the small differences in the variance component estimates observed among the three doubled-dataset parental models in Table 2.

There would likely also be some small differences in the variance component estimates in the doubled-dataset/parent models compared to *ASR_ind* and *ASR_par*, due to the artificial doubling of the data. As noted above, for the additive and additive $\times$ environment interaction terms, the number of females (and the number of males) in the doubled-dataset analysis equals the number of parents in *ASR_ind* or *ASR_par*, so we expect no bias in the variance component estimates associated with these random effects. However, for the random effects *female* $\times$ *male*, *female* $\times$ *male* $\times$ *environment*, and *residual*, we expect there to be a very small downward bias in the variance estimate compared to the true underlying variance. For a simple random variable $y \sim N(\mu, \sigma^2)$, a sample variance estimator s^2

$$s^2 = \frac{\sum_1^n (y - \bar{y})^2}{n - 1}$$

is an unbiased estimator of σ^2

$$E(s^2) = \sigma^2$$

It can be demonstrated that with a doubled dataset, the sample variance estimator s_d^2

$$s_d^2 = \frac{\sum_1^2 \sum_1^n (y - \bar{y})^2}{2n - 1}$$

is a biased estimator of σ^2 (see Appendix A).

$$E(s_d^2) = \left[\frac{2(n-1)}{2n-1} \right] \sigma^2$$

However, for mating designs of any size at all (i.e., any n of a moderate size, e.g., greater than 30), the effect of the bias should be negligible.

Once a mixed-model software package converges to a REML solution for the variance components associated with the random effects, the estimated variance components are then used to properly weight the observed data in the calculation of the solutions of random effects. The doubled-dataset/parent models gave essentially the same variance component estimates and genetic parameter estimates (functions of the variance components associated with the random terms in the models). There were some small differences among the software packages and/or linear models, but any differences had no important effect on the solutions of the random effects (breeding values and SCA predictions), as mentioned above.

It is important to note that any inferences regarding fixed effects (e.g., F-tests for differences among sites) would likely be affected by the doubling of the apparent number of degrees of freedom for many of the random effects. If such inferences are important, the researcher should be careful to test for these using a separate analysis with a non-doubled dataset.

4.3. Precision of Variance Component Estimates

It is also possible to demonstrate that the use of a doubled dataset of a random sample of y will produce a very small downward bias in the estimates of the precision for the variance component estimates, but again, this downward bias is negligible (see Appendix A). In other words, the expected value of the standard error estimate with a doubled dataset is approximately equal to the standard error estimate of the normal dataset:

$$SE(s_d^2) = SE(s^2) \left[\frac{2(n-1)}{2n-1} \right]$$

$$SE(s_d^2) \approx SE(s^2)$$

However, the use of the doubled-dataset/parent model algorithm for mixed-model analysis produces notably lower standard error estimates from the software output than is correct for some of the random effects (*female* $\times$ *male*, *female* $\times$ *male* $\times$ *environment*, and *residual*). This is due to a doubling of the apparent degrees of freedom, from $(n-1)$ to $(2n-1)$, and thus, we suggest that a correction factor $N = \sqrt{(2n-1)/(n-1)}$, where n = the degrees of freedom, should adjust the estimated standard errors. As n increases, the factor N approaches $\sqrt{2}$, so the general use of $N = \sqrt{2}$ should be a very good approximation. For example, for the *female* $\times$ *male* variance, multiplying the estimated standard errors from the doubled-dataset analyses by $N = \sqrt{2}$ gives an adjusted $se(\hat{\sigma}_{fm}^2) = 18.1$ and 18.2 from *ASR_parDouble* and *SAS_parDouble*, respectively, similar to the $se(\hat{\sigma}_{fm}^2) = 18.3$ from *ASR_ind*. Similarly, for the *female* $\times$ *male* $\times$ *environment* variance, multiplying the estimated standard errors from the doubled dataset analyses by the correction factor $N = \sqrt{2}$ gives an adjusted $se(\hat{\sigma}_{fm}^2) = 19.2$ and 19.4 from *ASR_parDouble* and *SAS_parDouble*, respectively, similar to the $se(\hat{\sigma}_{fm}^2) = 19.2$ from *ASR_ind*.

The confidence interval for the *female* $\times$ *male* variance arising from the *R_parDouble* analysis output also underestimates the range (or overestimates the precision) due to the apparent doubling of the *female* $\times$ *male* degrees of freedom. We suggest that an approximate adjustment of the range can be made

using the same adjustment factor N mentioned above. For example, for the *female x male* variance, $R_parDouble$ estimates $\hat{\sigma}_{fm}^2 = 35.6$, with an asymmetric 68.3% confidence interval from 23.8 to 49.8 ($\hat{\sigma}_{fm}^2 - 11.9$ to $\hat{\sigma}_{fm}^2 + 14.1$). Multiplying the lower and upper deviations by $N = \sqrt{2}$ gives an adjusted range of $\hat{\sigma}_{fm}^2 - 16.8$ to $\hat{\sigma}_{fm}^2 + 19.9$. This is roughly comparable to the range of the estimated $\hat{\sigma}_A^2 = 36.5 \pm 18.3$ from the ASR_ind model (Table 2). We believe that the proposed method for adjusting the standard error estimates of the variance components will be a very good approximation; however, the accuracy and robustness of the method should be confirmed by theoretical derivation or simulation studies.

4.4. Genetic Parameter Estimates

Breeders might also use genetic parameter estimates to guide breeding or testing strategies. For example, a breeder might want to examine how the number of progeny per site and the number of test sites would affect the precision of the breeding value predictions for the parents or progeny. Alternatively, a breeder might want to understand the size of the SCA variance relative to the additive genetic variance to determine if the marginal gain from the deployment of the best full-sib families would justify the cost of full-sib family testing versus half-sib family testing. Since the individual tree model and the doubled-dataset/parent models give the same genetic parameter estimates, the breeder will come to the same conclusions regarding breeding strategy alternatives.

4.5. Calculation of Individual Tree-Breeding Values with Half-Sib Progeny Test Data

This manuscript has demonstrated an approach to calculating the individual tree-breeding values of progeny using a doubled full-sib dataset and linear mixed model with *female* and *male* parents as random effects (see supplementary materials, Archive S1). The individual tree Mendelian deviation is calculated from the individual tree residual multiplied by the within-family heritability, and this is added to the mid-parent breeding value to calculate the individual tree-breeding value. The same basic approach will work with half-sib progeny test data with an appropriate adjustment of the linear model and the software code. Note that with a half-sib model, only one parent (typically the female or seed parent with forest trees) is included in the linear model, so there is no need to double the dataset. It is also important to note that the numerator of the within-family heritability for the half-sib case represents $\frac{3}{4}$ of the additive variance, rather than $\frac{1}{2}$ of the additive variance, as in the full-sib case.

4.6. Limitations: Multiple Generations

The doubled-dataset/parent model produces appropriate results with a single-generation full-sib dataset, i.e., a dataset with one distinct generation of parents and one distinct generation of progeny. For datasets with multiple generations, where some genotypes occur in the dataset as both parents and progeny, a parent model will not work. Multiple-generation datasets must be analyzed with software that allows for the pedigree function to calculate an appropriate numerator relationship matrix. For some organisms that typically produce small numbers of progeny (e.g., those in many animal-breeding programs), this will be a significant limitation. For some organisms that produce many progeny, this limitation is less problematic. For example, with forest trees (and many other plants), it is common for the parents of any given generation g to be represented by hundreds of progeny tested across a number of environments in replicated designs. Thus, each parental breeding value will be predicted quite accurately. Any progeny selected to be the parents of generation $g + 1$ will likewise be represented by many progeny, so those data can be analyzed as a stand-alone dataset to provide accurate breeding value predictions for generation $g + 1$'s parents.

4.7. Datasets with Both Full-Sib and Half-Sib Observations

Sometimes, a breeder might want to analyze a dataset that contains a mix of full-sib and half-sib progeny test data. Half-sib families (or open-pollinated families, derived from random mating in a seed orchard, native stand, or commercial planting) are easy to collect and, assuming a large population of

male (pollen) parents, can provide a good measure of the maternal breeding value. With a full-sib progeny dataset, each individual tree observation would be associated with two parental random effects, one for the female parent and one for the male parent. With half-sib data, each individual tree observation would be associated with only parental random effects, typically for the female parent. This is theoretically easy with a pedigree function that creates a Z design matrix for parents: a full-sib observation would have values in two columns, and a half-sib observation would have values in only one column. However, general linear mixed-model software packages do not allow a missing value for a random-effect classification variable (the male parent, in this case), and typically, observations with missing class variables are deleted from the dataset. In the case of a mixed full-sib and half-sib dataset, one option might be to code all the half-sib progeny as having a common unknown male parent. This might be appropriate if all of the half-sib families were collected from a common seed orchard. Another option might be to code each half-sib family with a unique unknown male parent. This might be a safer and more robust approach, which could account for different compositions of male parents, e.g., due to different seed collection sites, differences in flowering phenology, etc. If the proportion of half-sib progeny in the dataset is small, we believe there would be minimal effect on the variance component and genetic parameter estimates, but this is an area deserving further research, and breeders should exercise some caution.

4.8. Heterogeneous Variance Components

The doubled-dataset/parent model algorithm outlined here has been presented with a homogeneous variance example; that is, each random effect is assumed to have a single variance parameter that applies to the entire population of observations. This assumption may not always be valid, and a breeder might wish to model some heterogeneity in some of the random effects. For example, different test environments might have different residual variances depending on the inherent uniformity of the field site, or on differences in management. Another example could be tests of different classes or types of environments that produce differences in growth or different levels of genetic expression of a trait. If these situations exist, a breeder may want to include heterogeneity in the linear model. Whatever flexibility exists in a particular software program to allow for complex heterogeneous variance components in the **G** matrix or the **R** matrix could easily be used with the algorithm presented here.

5. Conclusions

Breeders who need to analyze a full-sib progeny test dataset, and wish to do so using a general linear mixed-model software package lacking a pedigree function, can accomplish this analysis using a doubled dataset and a model including *female* and *male* parents as random effects. This approach will provide variance component estimates, genetic parameter estimates, breeding value predictions for both the parents and progeny, and solutions to all other important random effects that are, for all practical purposes, identical to the results that would be obtained from an analysis using individual tree data and a pedigree function.

Supplementary Materials: The following archive is available online at <http://www.mdpi.com/1999-4907/11/11/1169/s1>. Archive S1: Algorithm FS MM.zip: This is a zip file with five folders, one for each software/linear model described here: *ASR_ind*, *ASR_parent*, *ASR_parDouble*, *SAS_parDouble*, and *R_parDouble*. Each folder contains one file with the code to run the model, and an “Output” subfolder with relevant files of parameter estimates and BLUP solutions. All files are either in text or .CSV format.

Author Contributions: Conceptualization, G.R.H.; Methodology, G.R.H. and J.J.A.; Software, G.R.H. and J.J.A.; Validation, G.R.H. and J.J.A.; Formal Analysis, G.R.H. and J.J.A.; Investigation, G.R.H. and J.J.A.; Resources, G.R.H. and J.J.A.; Data Curation, J.J.A.; Writing – Original Draft Preparation, G.R.H.; Writing – Review & Editing, G.R.H. and J.J.A.; Visualization, G.R.H.; Supervision, G.R.H.; Project Administration, J.J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Thanks go to Luis Ibarra for assistance with generating the simulated dataset, Fikret Isik for helpful tips on coding ASREML, and Salvador Gezan for discussions about the precision of the variance component estimates. Additionally, many thanks go to Camcore members for their long-term commitment to the tree improvement of tropical and sub-tropical pines and eucalypts.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Downward Bias in Estimates of σ^2 and $se(\sigma^2)$ When Using a Doubled Sample of a Random Variable

Consider a random variable $y \sim N(\mu, \sigma^2)$.

For a random variable Q , defined as

$$Q = \frac{\sum_1^n (y - \bar{y})^2}{\sigma^2}$$

Q has distribution $Q \sim \chi_{n-1}^2$ (based on Rencher and Schaalje [34], Theorem 5.5)

Therefore, the expected value and variance of Q are:

$$\begin{aligned} E(Q) &= (n-1) \\ Var(Q) &= 2(n-1) \end{aligned}$$

Now define a parameter s^2

$$s^2 = \frac{\sum_1^n (y - \bar{y})^2}{n-1} = \frac{\sigma^2 Q}{n-1}$$

$$\begin{aligned} E(s^2) &= E\left[\frac{\sigma^2 Q}{n-1}\right] \\ &= \frac{\sigma^2}{n-1} E(Q) \\ &= \frac{\sigma^2}{n-1} (n-1) \\ &= \sigma^2 \end{aligned}$$

$$\begin{aligned} Var(s^2) &= Var\left[\frac{\sigma^2 Q}{n-1}\right] \\ &= \left[\frac{\sigma^2}{n-1}\right]^2 Var(Q) \\ &= \left[\frac{\sigma^2}{n-1}\right]^2 2(n-1) \end{aligned}$$

$$\begin{aligned} SE(s^2) &= \left[\frac{\sigma^2}{n-1}\right] \sqrt{2} \sqrt{n-1} \\ &= \frac{\sqrt{2} \sigma^2}{\sqrt{n-1}} \end{aligned}$$

Now define a parameter s_d^2 , a variance estimator with a doubled dataset.

$$s_d^2 = \frac{\sum_1^2 \sum_1^n (y - \bar{y})^2}{2n-1} = \frac{2 \sum_1^n (y - \bar{y})^2}{2n-1} = \frac{2 \sum_1^n (y - \bar{y})^2 \sigma^2}{(2n-1) \sigma^2} = \frac{2 \sigma^2 Q}{2n-1}$$

$$\begin{aligned} E(s_d^2) &= E\left[\frac{2 \sigma^2 Q}{2n-1}\right] \\ &= \left[\frac{2 \sigma^2}{2n-1}\right] E(Q) \\ &= \left[\frac{2 \sigma^2}{2n-1}\right] (n-1) \\ &= \left[\frac{2(n-1)}{2n-1}\right] \sigma^2 \end{aligned}$$

$$\begin{aligned} Var(s_d^2) &= Var\left[\frac{2 \sigma^2 Q}{2n-1}\right] \\ &= \left[\frac{2 \sigma^2}{2n-1}\right]^2 Var(Q) \\ &= \left[\frac{2 \sigma^2}{2n-1}\right]^2 2(n-1) \end{aligned}$$

$$\begin{aligned}
 SE(s_d^2) &= \left[\frac{2\sigma^2}{2n-1} \right] \sqrt{2(n-1)} \\
 &= \sqrt{2} \sigma^2 \left[\frac{2}{2n-1} \right] \sqrt{(n-1)} \\
 &= \frac{\sqrt{2} \sigma^2}{\sqrt{(n-1)}} \left[\frac{2(n-1)}{2n-1} \right] \\
 &= SE(s^2) \left[\frac{2(n-1)}{2n-1} \right]
 \end{aligned}$$

References

- Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **1975**, *31*, 423–447. [\[CrossRef\]](#) [\[PubMed\]](#)
- Henderson, C.R. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics* **1976**, *32*, 69–83. [\[CrossRef\]](#)
- Kennedy, B.W.; Schaeffer, L.R.; Sorensen, D.A. Genetic Properties of Animal Models. *J. Dairy Sci.* **1988**, *71*, 17–26. [\[CrossRef\]](#)
- Isik, F.; Holland, J.; Maltecca, C. *Genetic Data Analysis for Plant and Animal Breeding*; Springer International Publishing: New York, NY, USA, 2017. [\[CrossRef\]](#)
- White, T.L.; Neale, D.; Adams, W.T. *Forest Genetics*; CABI Publishing: Wallingford, UK, 2007.
- Cornelius, J. Heritabilities and Additive Genetic Coefficients of Variation in Forest Trees. *Can. J. For. Res.* **1994**, *24*, 372–379. [\[CrossRef\]](#)
- Falconer, D.; Mackay, T. *Introduction to Quantitative Genetics*; Prentice Hall: Harlow, UK, 1996.
- Mrode, R.A. *Linear Models for the Prediction of Animal Breeding Values*; CABI Publishing: Wallingford, UK, 2014.
- Piepho, H.P.; Möhring, J.; Melchinger, A.E.; Büchse, A. BLUP for Phenotypic Selection in Plant Breeding and Variety Testing. *Euphytica* **2008**, *161*, 209–228. [\[CrossRef\]](#)
- Gilmour, A.; Gogel, B.; Cullis, B.; Welham, S.; Thompson, R. *ASREML User Guide Release 4.1*; VSN International: Hemel Hempstead, UK, 2015.
- Gilmour, A.R. Echidna Mixed Model Software. In Proceedings of the World Congress on Genetics Applied to Livestock Production, Volume Methods and Tools-Software, Auckland, New Zealand, 11–16 February 2018; pp. 995–998. Available online: <http://www.wcgalp.org/proceedings/2018/echidna-mixed-models-software> (accessed on 2 March 2020).
- Boldman, K.G.; Kriese, L.A.; Van Vleck, L.D.; Van Tassell, C.P.; Kachman, S.D. *A Manual for Use of MTDFREML—A Set of Programs to Obtain Estimates of Variances and Covariances*; United States Department of Agriculture, Agricultural Research Service, USDA, ASR: Beltsville, MD, USA, 1995.
- Resende, M.D.V. Software Selegen-REML/BLUP: A Useful Tool for Plant Breeding. *Crop Breed. Appl. Biotechnol.* **2016**, *16*, 330–339. [\[CrossRef\]](#)
- Meyer, K. WOMBAT: A Tool for Mixed Model Analyses in Quantitative Genetics by Restricted Maximum Likelihood (REML). *J. Zhejiang Univ. Sci. B* **2007**, *8*, 815–821. [\[CrossRef\]](#) [\[PubMed\]](#)
- SAS Institute Inc. *SAS/STAT 14.1 User's Guide*; SAS Institute Inc.: Cary, NC, USA, 2015.
- R Core Team. *R: A Language and Environment for Statistical Computing. R Version 4.0.2 (2020-06-22)–“Taking Off Again”*; The R Foundation for Statistical Computing: Vienna, Austria, 2020.
- Bates, D.; Mächler, M.; Bolker, B.M.; Walker, S.C. Fitting Linear Mixed-effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*. [\[CrossRef\]](#)
- Quaas, R.L.; Pollak, E.J. Mixed Model Methodology for Farm and Ranch Beef Cattle Testing Programs. *J. Anim. Sci.* **1980**, *51*, 1277–1287. [\[CrossRef\]](#)
- Burdon, R.D. Genetic Correlation as a Concept for Studying Genotype-environment Interaction in Forest Tree Breeding. *Silvae Genet.* **1977**, *26*, 168–175.
- Dickerson, G. *Techniques and Procedures in Animal Science Research*; American Society of Animal Science: Champaign, IL, USA, 1969.
- Lee, E.S.; Forthofer, R.N. *Analyzing Complex Survey Data*; SAGE Publications: Thousand Oaks, CA, USA, 2006.
- Bates, D.M. Assessing the Precision of Estimates of Variance Components. Presentation at Computacionale Statistik, Ludwig Maximilian University, Munich, Germany, 16 July 2009. Available online: <http://lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4PrecisionD.pdf> (accessed on 19 May 2020).

23. Hodge, G.R.; Dvorak, W.S. Growth Potential and Genetic Parameters of Four Mesoamerican Pines Planted in the Southern Hemisphere. *South. For. J. For. Sci.* **2012**, *74*, 27–49. [[CrossRef](#)]
24. Lstibůrek, M.; Hodge, G.R.; Lachout, P. Uncovering Genetic Information from Commercial Forest Plantations—Making Up for Lost Time Using “Breeding without Breeding”. *Tree Genet. Genomes* **2015**, *11*. [[CrossRef](#)]
25. White, T.L.; Hodge, G. *Predicting Breeding Values with Applications in Forest Tree Improvement*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989.
26. Searle, S.R.; Casella, G.; McCulloch, C.E. *Variance Components*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2009.
27. Zobel, B.; Talbert, J. *Applied Forest Tree Improvement*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1984.
28. Lstibůrek, M.; El-Kassaby, Y. Advanced-Generation Seed Orchard Designs. In *Seed Orchards, Proceedings from a Conference at Umeå, Sweden, 26–28 September 2007*; IUFRO International Union of Forest Research Organizations: Vienna, Austria, 2008; pp. 155–161.
29. McKeand, S.; Jett, J.; O’Berry, S.; Heine, A. New Challenges for Seed Orchard Management of Loblolly Pine in the Southern US. In *Proceedings of the IUFRO Seed Orchard Conference 2017, Bålsta, Sweden, 4–6 September 2017*; Swedish University of Agricultural Sciences: Umeå, Sweden, 2017.
30. Johnson, D.L.; Thompson, R. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. *J. Dairy Sci.* **1995**, *78*, 449–456. [[CrossRef](#)]
31. Gilmour, A.R.; Thompson, R.; Cullis, B.R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* **1995**, *51*, 1440–1450. [[CrossRef](#)]
32. Knight, E. Improved Iterative Schemes for REML Estimation of Variance Parameters in Linear Mixed Models. Ph.D. Thesis, The University of Adelaide, Adelaide, Australia, 2008.
33. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
34. Rencher, A.; Schaalje, G. *Linear Models in Statistics*, 2nd ed.; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2008; pp. 117–118.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).