Ordinary Least Squares Estimation of General and Specific Combining Abilities from Half-Diallel Mating Designs[1])

By D. A. Huber[a]), T. L. White[a]), R. C. Littell[b]) and G. R. Hodge[a])

Department of Forestry[a]),
Department of Statistics[b]),
University of Florida,
Gainesville, FL 32611, USA

(Received 16th December 1991)

Abstract

This paper presents a matrix algebra approach to the solution of ordinary least squares (OLS) equations for estimation of genetic parameters as fixed effects in a half-diallel mating system. The relative ease of implementation and omnipresence of OLS based analysis in forest genetics applications motivates the current discussion. Few statistical packages are available that allow direct specification of a half-diallel model and dependence on standard formulae for estimation can lead to erroneous results if the data are unbalanced (plots or crosses are missing). The methods offered herein can be implemented with any software capable of matrix manipulations and can be extended to any linear model.

The mechanics of the OLS analysis are described to foster understanding of the assumptions and mathematical properties of ordinary least squares analysis and sum-to-zero restrictions. The impacts of OLS assumptions and sum-to-zero restrictions on the interpretation of genetic parameter estimates are examined, and the relationship between OLS assumptions and other types of analyses is presented. The analysis is developed through formation of the design matrix by modeled parameters as an extension of the scalar linear model for half-diallel designs. The overparameterized design matrix is created for balanced data (plot-mean basis) and is then reparameterized to full-rank in sum-to-zero format. Adjustments to the design matrix prompted by various common causes of data imbalance are demonstrated. Numerical examples are provided along with the data set so that the reader may recreate the analysis.

Two problems which are basic in a genetic analysis are addressed: 1) the variance-covariance matrix for the observations, and 2) the desire to compare breeding value estimates from disconnected experiments. In general, OLS assumptions about the variance-covariance matrix for cell (plot) means are not appropriate and evaluation of these assumptions often suggests other types of analyses especially for large data bases. Comparison of breeding value estimates from disconnected experiments is problematic with any analysis method; however, a method is offered to improve OLS estimate comparability.

Key words: Unbalanced data, estimability.

Introduction

The diallel mating system is an altered factorial design in which the same individuals (or lines) are used as both male and female parents. A full diallel contains all crosses, including reciprocal crosses and selfs, resulting in a total of p^2 combinations, where p is the number of parents. Assumptions that reciprocal effects, maternal effects, and paternal effects are negligible lead to the use of the half-diallel mating system (Griffing, 1956, method 4) which has p(p—1)/2 parental combinations and is the mating system addressed in this paper.

Half diallels have been widely used in crop and tree breeding (Sprague and Tatum, 1942; Gilbert, 1958; Matzinger et al., 1959; Burley et al., 1966; and Squillace, 1973) and the widespread use of this mating system continues today (Weir and Zobel, 1975; Wilcox et al., 1975; Snyder and Namkoong, 1978; Hallauer and Miranda, 1981; Singh and Singh, 1984; Greenwood et al., 1986; and Weir and Goddard, 1986).

Most of the statistical packages available treat fixed effect estimation as the objective of the program with random variables representing nuisance variation. Within this context a common analysis of half-diallel experiments is conducted by first treating genetic parameters as fixed effects for estimation of general (GCA) and specific (SCA) combining abilities and subsequently as random variables for variance component estimation (used for estimating heritabilities, genetic correlations, and general to specific combining ability variance ratios for determining breeding strategies). This paper focuses on the estimation of GCA's and SCA's as fixed effects. The treatment of GCA and SCA as fixed effects in OLS (ordinary least squares) is an entirely appropriate analysis if the comparisons are among parents and crosses in a particular experiment. If, as forest geneticists often wish to do, GCA estimates from disconnected experiments are to be compared, then methods such as checklots must be used to place the estimates on a common basis.

Formulae (Griffing, 1956; Falconer, 1981; Hallauer and Miranda, 1981; and Becker, 1975) for hand calculation of general and specific combining abilities are based on a solution to the OLS equations for half-diallels created by sum-to-zero restrictions i. e. the sum of all effect estimates for an experimental factor equals zero). These formulae will yield correct OLS solutions for sum-to-zero genetic parameters provided the data have no missing cells. If cell (plot) means are used as the basis for the estimation of effects there must be at least one observation per cell (plot) where a cell is a subclassification of the data defined by one level of every factor (Searle, 1987). An example of a cell is the group of observations denoted by AB_{ij} for a randomized complete block design with factor A across blocks (B). If the above formulae are applied without accounting for missing cells, incorrect and possibly misleading solutions can result. The matrix algebra approach is described in this paper for these reasons: 1) in forest tree breeding applications data sets with missing cells are extremely common; 2) many statistical packages do not allow direct specification of the half-

[1]) This is Journal Series No. R-02358 of the Institute of Food and Agricultural Sciences.

diallel model; 3) the use of a linear model and matrix algebra can yield relevant OLS solutions for any degree of data imbalance; and 4) viewing the mechanics of the OLS approach is an aid to understanding the properties of the estimates.

The objectives of this paper are to: (1) detail the construction of ordinary least squares (OLS) analysis of half-diallel data sets to estimate genetic parameters (GCA and SCA) as fixed effects, (2) recount the assumptions and mathematical features of this type of analysis, (3) facilitate the reader's implementation of OLS analyses for diallels of any degree of imbalance and suggest a method for combining estimates from disconnected experiments, and (4) aid the reader in ascertaining what method is an appropriate analysis for a given data set.

Methods

2.1 Linear Model

Plot means are used as the unit of observation for this analysis with unequal numbers of observations per plot. Plot (cell) means are always estimable as long as there is one observation per plot, and linear combinations of these means (least squares means) provide the most efficient way of estimating OLS fixed effects (YATES, 1934). Throughout this paper, estimates are denoted by lower case letters while the parameters are designated by upper case letters and matrices are in bold print.

Using plot means as observations, a common scalar linear model for an analysis of a half-diallel mating design with p(p—1)/2 crosses planted at a single location in a randomized complete block design with one plot per block is:

$$y_{ijk} = \mu + B_i + GCA_j + GCA_k + SCA_{jk} + e_{ijk} \qquad (1)$$

where: y_{ijk} = the mean of the i^{th} block for the jk^{th} cross
μ = an overall mean
B_i = fixed effect of block i for i = 1 to b;
GCA_j = fixed general combining ability effect of the j^{th} female parent or k^{th} male parent, j or k = 1,. . .,p ($j \neq k$);
SCA_{jk} = fixed specific combining ability effect of parents j and k; and
e_{ijk} = the random error associated with the observation of the jk^{th} cross in the i^{th} block, $e_{ijk} \sim (0, \sigma^2_e)$.

Cross by block interaction as genotype by environment interaction is treated as confounded with between plot variation as for contiguous plots.
The model in matrix notation is:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad (2)$$

where: y is the vector of observation vectors (nx1 = n rows and 1 column) where n equals the number of observations;
X is the design matrix (nxm) whose function is to select the appropriate parameters for each observation where m equals the number of fixed effect parameters in the model;
β is the vector (mx1) of fixed effect parameters ordered in a column; and
e is the vector (nx1) of deviations (errors) from the expectation associated with each observation.

2.1.1 Ordinary Least Squares Solutions

The matrix representation of an OLS fixed effects solution is:

$$\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad (3)$$

where: $\hat{b}$ is the vector of estimated fixed effect parameters, i.e. an estimate of β,
X is the design matrix full rank (by reparameterization),
or a generalized inverse of X'X may be used,
and y is the vector of observations.
where j for GCA is j or k and τ represents the fixed genetic parameter of the checklot.

Inherent in this solution is the ordinary least squares assumption that the variance-covariance matrix (V) of the observations (y) is equal to $I\sigma^2_e$, where I is an nxn identity matrix. The elements of an identity matrix are 1's on the main diagonal and all other elements are 0. Multiplying I by σ^2_e places σ^2_e on the main diagonal. In the variance-covariance matrix for the observations, the variance of the observations appears on the main diagonal and the covariance between observations appears in the off-diagonal elements. Thus, $V = I\sigma^2_e$ states that the variance of the observations is equal to σ^2_e for each observation and there are no covariances between the observations (which is one direct result of considering genetic parameters as fixed effects).

2.1.2 Sum-to-Zero Restrictions

The design matrix presented in this paper is reparameterized by sum-to-zero restrictions to: (1) reduce the dimension of the matrices to a minimal size, and (2) yield estimates of fixed effects with the same solution as common formulae in the balanced case. Other restrictions such as set-to-zero could also be applied so the discussion that follows treats sum-to-zero restrictions as a specific solution to the more general problem which is finding an inverse for X'X. The subscripts 'o' and 's' refer to the overparameterized model and the reparameterized model with sum-to-zero restrictions, respectively.

The matrix X_0 of *figure 1* is the design matrix for an overparameterized linear model (MILLIKEN and JOHNSON, 1984, page 96). Overparameterization means that the equations are written in more unknowns (parameters, in this case 13) than there are equations (number of observations minus degrees of freedom for error, in this case 12 — 5 = 7) with which to estimate the parameters. Reparameterization as a sum-to-zero matrix overcomes this dilemma by reducing the number of parameters through making some of the parameters linear combinations of others. Sum-to-zero restrictions make the resulting parameters and *estimates* sum to zero even though the unrestricted parameters (for example the true GCA values as applied to a broader population) do not necessarily sum-to-zero within a diallel. This is the problem of comparability of GCA estimates from disconnected experiments.

To illustrate the concept of sum-to-zero estimates versus population parameters, we use the expectation of a common formula. BECKER (1975) gives *equation 4* (which for balanced cases is equivalent to $g_j = ((p-1)/(p-2))(\bar{Z}_{j.} - \bar{Z}_{..})$) as the estimate for general combining ability for the j^{th}

line with p equalling the number of parents and Z_{jk} equalling the site mean of the j x k cross. This equation yields the same solution as the matrix equations with no missing plots or crosses and with a design matrix which contains the sum-to zero restrictions. An evaluation of this formula in a four-parent half-diallel planted in b blocks for the GCA of parent 1 is obtained by substituting the expectation of the linear model (*Equation 1*) for each observation:

$$g_j = (1/(p(p-2)))(pZ_{j.} - 2Z_{..}) \tag{4}$$

$$E\{g_1\} = E\{(1/(p(p-2)))(pZ_{1.} - 2Z_{..})\}$$

$$E\{g_1\} = 3/4(GCA_1) - 1/4(GCA_2 + GCA_3 + GCA_4) + 1/4(SCA_{12} + SCA_{13} + SCA_{14}) -$$

$$1/4(SCA_{23} + SCA_{24} + SCA_{34}).$$

The result of *equation 4* is obviously not GCA_1 from the unrestricted model (*Equation 1*). Thus, g_1, an estimable function and an estimate of parameter GCA_{1s} (the estimate of the GCA of parent 1 given the sum-to-zero restructions), does not have the same meaning as GCA_1 in the unrestricted model. An estimable function is a linear combination of the observations; but in order for an individual parameter in a model to be estimable, one must devise a linear combination of the observations such that the expectation has a weight of one on the parameter one wishes to estimate while having a weight of zero on all other parameters. A solution such as this does not exist for the individual parameters in the overparameterized model (*Equation 1*). So although the sum-to-zero restricted GCA parameters and estimates are forced to sum-to-zero for the sample of parents in a given diallel, the unrestricted GCA parameters only sum-to-zero across the entire population (FALCONER, 1981) and an evaluation of GCA_{1s} demonstrates that the estimate contains other model parameters.

The result of sum-to-zero restrictions is that the degrees of freedom for a factor equals the number of columns (parameters) for that factor in X_s (*Figure 2*). Thus, a generalized inverse for $X_s'X_s$ is not required since the number of columns in the sum-to-zero X_s matrix for each factor equals the degrees of freedom for that factor in the model (X_s is full column rank and provides a solution to *Equation 3).*

$$\begin{bmatrix} y_{112}\\ y_{113}\\ y_{114}\\ y_{123}\\ y_{124}\\ y_{134}\\ y_{212}\\ y_{213}\\ y_{214}\\ y_{223}\\ y_{224}\\ y_{234} \end{bmatrix} = \begin{array}{c} \begin{array}{ccccccccccccc} \mu & B_1 & B_2 & GCA_1 & GCA_2 & GCA_3 & GCA_4 & SCA_{12} & SCA_{13} & SCA_{14} & SCA_{23} & SCA_{24} & SCA_{34} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0\\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0\\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1\\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0\\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0\\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{bmatrix} \mu\\ B_1\\ B_2\\ GCA_1\\ GCA_2\\ GCA_3\\ GCA_4\\ SCA_{12}\\ SCA_{13}\\ SCA_{14}\\ SCA_{23}\\ SCA_{24}\\ SCA_{34} \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}_o \boldsymbol{\beta}_o$$

Figure 1. — The overparameterized linear model for a four-parent half-diallel planted on a single site in two blocks displayed as matrices. The design matrix(X_0) and parameter vector (β_0) are shown in overparameterized form. 1's and O's denote the presence or absence of a parameter in the model for the observed means (data vector, y). The parameters displayed above the design matrix label the appropriate column for each parameter. Error vector not exhibited.

$$\begin{bmatrix} y_{112}\\ y_{113}\\ y_{114}\\ y_{123}\\ y_{124}\\ y_{134}\\ y_{212}\\ y_{213}\\ y_{214}\\ y_{223}\\ y_{224}\\ y_{234} \end{bmatrix} = \begin{array}{c} \begin{array}{ccccccc} \mu & B_1 & GCA_1 & GCA_2 & GCA_3 & SCA_{12} & SCA_{13} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0\\ 1 & 1 & 1 & 0 & 1 & 0 & 1\\ 1 & 1 & 0 & -1 & -1 & -1 & -1\\ 1 & 1 & 0 & 1 & 1 & -1 & -1\\ 1 & 1 & -1 & 0 & -1 & 0 & 1\\ 1 & 1 & -1 & -1 & 0 & 1 & 0\\ 1 & -1 & 1 & 1 & 0 & 1 & 0\\ 1 & -1 & 1 & 0 & 1 & 0 & 1\\ 1 & -1 & 0 & -1 & -1 & -1 & -1\\ 1 & -1 & 0 & 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 0 & -1 & 0 & 1\\ 1 & -1 & -1 & -1 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{bmatrix} \mu\\ B_1\\ GCA_1\\ GCA_2\\ GCA_3\\ SCA_{12}\\ SCA_{13} \end{bmatrix} + \begin{bmatrix} e112\\ e113\\ e114\\ e123\\ e124\\ e134\\ e212\\ e213\\ e214\\ e223\\ e224\\ e234 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}_s \boldsymbol{\beta}_s + \mathbf{e}.$$

Figure 2. — The linear model for a four-parent half-diallel planted on a single site in 2 blocks displayed as matrices. The design matrix (X_s) and the parameter vector (β_s) are presented in sum-to-zero format. The parameters displayed above the design matrix label the appropriate column for each parameter.

2.2 Components of the Matrix Equation

The equational components of 2 are now considered in greater detail.

2.2.1 Data Vector y

Observations (plot means) in the data vector are ordered in the manner demonstrated in *figure 1*. For our example *figure 1* is the matrix equation of a four parent half-diallel mating design planted in two randomized complete blocks on a single site. There are six crosses present in the two blocks for a total of 12 observations in the data vector, y. The observations are first sorted by block. Second, within each block the observations should be in the same sequence (for simplicity of presentation only). This sequence is obtained by assigning numbers 1 through p to each of the p parents and then sorting all crosses containing parent 1 (whether as male or female) as the primary index in descending numerical order by the other parent of the cross as the secondary index. Next all crosses containing parent 2 (primary index, as male or female) in which the other parent in the cross (secondary index) has a number greater than 2 are then also sorted in descending order by the secondary index. This procedure is followed through using parent p—1 as the primary index.

2.2.2 Design Matrix and Parameter Vector, X and β

The design matrix for a model is conceptually a listing of the parameters present in the model for each observation (SEARLE, 1987, page 243). In *figure 1*, y and β_0 are exhibited and the parameters in β_0 are displayed at the tops of the columns of X_0 (a visually correct interpretation of the multiplication of a matrix by a vector). For each observation in y, the scalar model (*Equation 1*) may be employed to obtain the listing of parameters for that observation (the row of the design matrix corresponding to the particular observation). The convention for design matrices is that the columns for the factors occur in the same order as the factors in the linear model (*Equation 1* and *Figure 1*). Since design matrices can be devised by first creating th columns pertinent to each factor in the model (submatrices) and then horizontally and/or vertically stacking the submatrices, the discussion of the reparameterized design matrix formulation will proceed by factor.

2.2.3 Mean

The first column of X_s is for μ and is a vector of l's with the number of rows equalling the number of observations (*Figure 2*). The linear model (*Equation 1*) indicates that all observations contain μ and the deviation of the observations from μ is explained in terms of the factors and interactions in the model plus error.

2.2.4 Block

The number of columns for block is equal to the number of blocks minus one (column 2, X_s). Each row of a block submatrix consists of 1's and 0's or —1's according to the identity of the observation for which the row is being formed. The normal convention is that the first column represents block 1 and the second column block 2, etc. through block b—1. Since we have used a sum-to-zero solution ($\Sigma_i b_i = 0$), the effect due to block b is a linear combination of the other b—1 effects, i. e. $b_b = -\Sigma_{i=1}^{b-1} b_i$ which in our example is $0 = b_1 + b_2$ and $b_2 = -b_1$. Thus, the row of the block submatrix for an observation in block (b the last block) has a —1 in each of the b—1 columns signifyng that the block b effect is indeed a linear combination of the other b—1 block effects. Columns 2 and 3 of X_0 (*Figure 1*) have become column 2 of X_s (*Figure 2*).

2.2.5 General Combining Ability

This submatrix of X_s is slightly more complex than previous factors as a result of having two levels of a main effect present per observation, i. e. the deviation of an observation from μ is modeled as the result of the GCA's of both the male and female parents (*Equation 1*). Again we have imposed a restriction, $\Sigma_j gca_j = 0$. Since GCA has p—1 degrees of freedom, the submatrix for GCA should have p—1 columns, i. e. $gca_p = -\Sigma_{j=1}^{p-1} gca_j$. The GCA submatrix for X_s (columns 3 through 5 in *Figure 2*) is formed from X_0 (columns 4 through 7 in *Figure 1*) according in the same manner as the block matrix: (1) add minus one to the elements in the other columns along each row containing a one for gca_p (p = 4 in our example); and (3) delete the column from X_0 corresponding to gca_p. The GCA submatrix has p(p—1)/2 rows (the number of crosses).This, with no missing cells (plots), equals the number of observations per block. To form the GCA factor submatrix for a site, the GCA submatrix is vertically concatenated (stacked on itself) b times. This completes the portion of the X_s matrix for GCA.

2.2.6 Specific Combining Ability

In order to facilitate construction of the SCA submatrix, a horizontal direct product should be defined. A horizontal direct product, as applied to two column vectors, is the element by element product between the two vectors (SAS/IML[2]) User's Guide, 1985) such that the element in the i<u>th</u> row of the resulting product vector is the product of the elements in the i<u>th</u> rows of the two initial vectors. The resultant product vector has dimension n x 1. A horizontal direct product is useful for the formation of interaction or nested factor submatrices where initial matrices represent the main factors and the resulting matrix represents an interaction or a nested factor (product rule, SEARLE, 1987).

The SCA submatrix can be formulated from the horizontal direct products of the columns of the GCA submatrix in X_s (*Figure 2*). The results from the GCA columns require manipulation to become the SCA submatrix (since degrees of freedom for SCA do not equal those of an interaction for a half-diallel analysis), but the GCA column products provide a convenient starting point. The column of the SCA submatrix representing the cross between the j^{th} and the k^{th} parents (SCA_{jk}) is formed as the product between the GCA_j and GCA_k columns (*Figure 3*). The GCA columns in *figure 2* are multiplied in this order: column 1 times column 2 forming the first SCA column, column 1 times column 3 forming the second SCA column, and column 2 times column 3 forming the third SCA column (*Figure 3*). With four parents (six crosses) there are three degrees of freedom for GCA (p—1) and two degrees of freedom for SCA (6 crosses — 3 for GCA — 1 for the mean). Since SCA has only two degrees of freedom, a sum-to-zero design matrix can have only two columns for SCA. Imposing the restriction that the sum of the SCA's across all parents equals zero is equivalent to making the last column for the SCA submatrix (*Figure*

[2]) SAS/IML is the registered trademark of the SAS Institute Inc. Cary, North Carolina.

OBS.	$GCA_1 x GCA_2$	$GCA_1 x GCA_3$	$GCA_2 x GCA_3$	SCA_{12}	SCA_{13}	SCA_{23}
Y_{i12}	(1)(1)=1	(1)(0)=0	(1)(0)=0	1	0	0
Y_{i13}	(1)(0)=0	(1)(1)=1	(0)(1)=0	0	1	0
Y_{i14}	(0)(-1)=0	(0)(-1)=0	(-1)(-1)=1	0	0	1
Y_{i23}	(0)(1)=0	(0)(1)=0	(1)(1)=1	0	0	1
Y_{i24}	(-1)(0)=0	(-1)(-1)=1	(0)(-1)=0	0	1	0
Y_{i34}	(-1)(-1)=1	(-1)(0)=0	(-1)(0)=0	1	0	0

Figure 3. — Intermediate result in SCA submatrix generation (SCA columns as horizontal direct products of GCA_1, GCA_2, and GCA_3 columns within a block). The SCA_{jk} column is the horizontal direct product of the columns for GCA_j and GCA_k.

3) a linear combination of the others *(Figure 2)*. The procedure for deleting the third column product is identical to that for the GCA submatrix: add minus one to every element in the rows of the remaining SCA columns in which a one appears in the column which is to be deleted (*Figure 2*, columns 6 and 7). The number of rows in the SCA submatrix equals the number observations in a block and must be vertically concatenated b times to create the SCA submatrix for a site.

An algebraic evaluation of SCA sum-to-zero restrictions requires that $\Sigma_j sca_{jk} = 0$ for each k and that $\Sigma_j \Sigma_k sca_{jk} = 0$; thus, for observations in the i^{th} block with i serving to denote the row of the SCA submatrix in block i, $sca_{i14} = -sca_{i12} - sca_{i13}$ and entries in the submatrix row for y_{i14} are —1's. Sca_{i23} equals sca_{i14} because sca_{i23} is the negative of the sum of the independently estimated SCA's (sca_{i12} and sca_{i13}) from the restriction that the sum of the SCA's across all parents equals zero. Similarly, by sum-to-zero definition $sca_{i24} = - sca_{i23} - sca_{i12}$ and by substitution $sca_{i24} = -(-sca_{i12} - sca_{i13}) - sca_{i12} = sca_{i13}$. By the same protocol, it can be shown that $sca_{i34} = sca_{i12}$. The elements in the rows of the SCA submatrix are 1's, —1's and 0's in accordance with the algebraic evaluation. Thus, while it may seem that there should be 6 SCA values (one for each cross), only 2 can be independently estimated and the remaining 4 are linear combinations of the independently estimated SCA's. Again the SCA sum-to-zero estimates are not equal to the parametric population SCA's. An analogous illustration for SCA to that for GCA would show that the estimable function (linear combination of observations) for a given SCA_s contains a variety of other parameters.

2.3 Estimation of Fixed Effects

2.3.1 GCA Parameters

The GCA parameters can be estimated (without mean, block, and SCA in the design matrix) through the use of *equation 3*, if there are no missing cell means (plots) for any cross and no missing crosses. The design matrix consists only of the GCA submatrix. This design matrix has (p—1) (for GCA's) columns (the third through the fifth columns of X_s). The b vector is an estimate of the GCA portion of β_s as in *figure 2* and the linear combinations for the estimation of gca_p is $gca_p = -\Sigma_{j=1}^{p-1} gca_j$. Parameters for any of the factors can be estimated independently using the pertinent submatrix as long as there are no missing cell means (plots) and no missing crosses; this uses a property known as orthogonality.

Orthogonality requires that the dot product between two vectors equals zero (SCHNEIDER, 1987, page 168). The dot product (a scalar) is the sum of the values in a vector obtained from the horizontal direct product of two vectors. For two factors to be orthogonal, the dot products of all the column vectors making up the section of the design matrix for one factor with the column vectors making up the portion of the design matrix for the second must be zero. If all factors in the model are orthogonal, then the $X_s'X_s$ matrix is block diagonal. A block-diagonal $X_s'X_s$ matrix is composed of square factor submatrices (degrees of freedom x degrees of freedom) along the diagonal with all offdiagonal elements not in one of the square factor submatrices equalling zero. A property of block-diagonal matrices is that the inverse can be calculated by inverting each block separately and replacing the original blocks in the full X'X matrix by the inverted block. Because the blocks can be inverted separately and all other off-diagonal elements of the inverse are zero, the effects for factors which are orthogonal to all other factors may be estimated separately. i. e. there are no functions of other sum-to-zero factors in the sum-to-zero estimates.

2.3.2 Mean, Block, GCA, and SCA Parameters

All parameters are estimated simultaneously by horizontally concatenating the mean, block, GCA, and SCA matrices to create X_s. *Equation 3* is again utilized to solve the system of equations. The b vector for the four parent example is an estimate of β_s of *figure 2*. Again, one parameter is estimated for each column in the X_s matrix and all parameter estimates not present are linear combinations of the parameter estimates in the b vector. So b_b is equal to $-\Sigma_{i=1}^{b-1} b_i$ and gca_p is equal to $-\Sigma_{j=1}^{p-1} gca_j$. The linear combinations for SCA effects can be obtained by reading along the row of the SCA submatrix associated with the observation containing the parameter, i. e. in *figure 2* the observation y_{i23} contains the effect sca_{i23} which is estimated as the linear combination $-sca_{i12} - sca_{i13}$.

This completes the estimation of fixed effect parameters from a data set which is balanced on a plot-mean basis. Since field data sets with such completeness are a rarity in forestry applications, the next step is OLS analysis for various types of data imbalance. Calculations of solutions based on a complete data set and simulated data sets with common types of imbalance are demonstrated in numerical examples.

Numerical Examples

The data set analyzed in the numerical examples is from a five-year-old, six-parent half-diallel slash pine (*Pinus elliottii* var. *elliotti* ENGELMN.) progeny test planted on a single site in four complete blocks. Each cross is represented by a five-tree row plot within each block. Total height in meters and diameter at breast height (dbh in centimeters) are the traits selected for analysis. The data set is presented in *table 1* so that the reader may reconstruct the analysis and compare answers with the examples. The numbers 1 through 6 were arbitrarily assigned to the parents for analysis. Because of unequal survival within plots, plot means are used as the unit of observation.

Table 1. — Data Set for Numerical Examples. Five-year-old slash pine progeny test with a 6-parent half-diallel mating design present on a single site with four randomized complete blocks and a five-tree row plot per cross per block.

Block	Female	Male	Mean Height	Mean DBH	Within Plot Variance Height	Within Plot Variance DBH	Trees per Plot
			Meters	*Centimeters*	*m²*	*cm²*	
1	1	2	2.6899	3.810	0.9800	3.484	4
1	1	3	1.9080	2.134	1.4277	3.893	5
1	1	5	3.1242	4.445	0.4487	1.656	4
1	1	6	2.4933	3.200	0.8488	5.664	5
1	2	5	1.4783	1.588	0.6556	2.167	4
1	2	6	2.7026	3.471	0.1136	0.344	3
1	3	2	3.0480	4.699	0.2341	0.968	4
1	3	5	3.4991	5.131	0.0945	0.271	5
1	3	6	2.4003	2.794	0.5149	1.548	4
1	4	1	3.3955	4.928	0.1489	0.761	5
1	4	2	3.4290	5.144	0.7943	3.285	4
1	4	3	2.5298	2.984	0.9557	4.188	4
1	4	5	2.4155	3.175	0.5936	2.946	4
1	4	6	3.2004	4.521	1.7034	7.594	5
1	5	6	2.2403	2.794	1.0433	6.280	4
2	1	2	3.5662	5.080	0.9560	2.903	5
2	1	3	2.6335	3.353	0.7695	3.497	5
2	1	5	3.6942	5.893	0.0573	0.432	5
2	1	6	3.4808	4.928	0.9222	2.890	5
2	2	5	3.4260	4.877	0.7017	2.432	5
2	2	6	2.4282	3.302	0.0616	0.452	3
2	3	2	3.0480	4.064	0.0192	0.301	4
2	3	5	2.8895	4.013	0.1957	0.690	5
2	3	6	1.9406	1.863	0.0560	0.408	3
2	4	1	3.0114	3.962	1.9753	6.342	5
2	4	2	3.6454	5.283	0.1731	0.787	5
2	4	3	2.9566	3.861	0.0506	0.174	5
2	4	5	2.8118	4.382	1.1336	5.435	4
2	4	6	3.2674	4.318	1.1211	4.354	5
2	5	6	3.7917	5.893	0.0848	0.497	5
3	1	2	2.2961	2.625	0.3914	1.699	3
3	1	3	2.8956	4.128	1.2926	4.532	4
3	1	5	2.5359	3.607	0.8284	4.303	5
3	1	6	2.9032	3.937	0.8252	4.064	4
3	2	5	2.7737	4.064	0.9829	3.226	2
3	2	6	1.2040	0.635	0.4464	0.806	2
3	3	2	2.9870	4.191	0.9049	2.989	4
3	3	5	2.8407	3.962	0.7309	3.632	5
3	3	6	1.3564	0.000	0.1677	0.000	2
3	4	1	2.6746	3.620	0.8463	2.984	4
3	4	2	2.7066	3.353	0.5590	1.787	5
3	4	3	3.4198	4.623	0.3509	0.690	5
3	4	5	3.3299	4.953	0.4102	1.226	4
3	4	6	3.4564	4.978	0.8369	3.503	5
3	5	6	3.2614	4.826	.	.	1
4	1	2	1.8974	2.476	1.0160	3.629	4
4	1	3	1.3005	0.508	0.2019	0.774	3
4	1	5	2.0726	2.540	1.2235	5.097	3
4	1	6	1.8821	1.778	0.4728	3.312	4
4	2	5	1. 64	1.334	0.5354	2.382	4
4	2	6	1.5392	0.635	0.0376	0.806	2
4	3	2	1.8898	2.032	0.7364	1.892	4
4	3	5	2.5146	3.620	0.0876	0.446	4
4	3	6	1.8389	2.201	0.0941	0.280	3
4	4	1	2.3348	2.591	0.3816	2.722	5
4	4	2	1.7272	1.693	2.1640	8.602	3
4	4	3	1.6581	1.524	0.0537	0.903	5
4	4	5	2.1184	2.286	0.3137	2.366	4
4	4	6	1.5545	1.422	0.4803	1.019	5
4	5	6	1.4122	1.693	0.0338	0.150	3

3.1 Balanced Data (Plot-mean Basis)

The sum-to-zero design matrix for the balanced data set has (4 blocks) x (15 crosses) = 60 rows (which equals the number of observations in y) and has the following columns: one column for μ, three columns for blocks (b—1), five columns for GCA (p—1), and nine columns for SCA (15 crosses — 5 — 1) for a total of 18 columns. With 60 plot means (degrees of freedom) and 18 degrees of freedom in the model, subtracting 18 from 60 yields 42 degrees of freedom for error which matches the degrees of freedom for cross by block interaction, thus verifying that degrees of fredom concur with the number of columns in the sum-to-zero design matrix.

To illustrate the principle of orthogonality in the balanced case, the $X_8'X_8$ and $(X_8'X_8)^{-1}$ matrices may be printed to show that they are block diagonal. In further illustration, the effects within a factor may also be estimated without any other factors in the design matrix and compared to the estimates from the full design matrix.

The vectors of parameter estimates for height and dbh (*table 2*) were calculated from the same X_8 matrix because height and dbh measurements were taken on the same trees. In other words, if a height measurement was taken on a tree, a dbh measurement was also taken, so the design matrices are equivalent.

Table 2. — Numerical results for examples of data imbalance using the OLS techniques presented in the text.

Estimate	Balanced[a]		Missing Plot[b]		Missing Cross[c]		Five Missing Crosses[d]	
of[e]	Height	DBH	Height	DBH	Height	DBH	Height	DBH
μ	2.5830	3.362	2.5787	3.346	2.5386	3.260	2.4980	3.149
B_1	0.1203	0.292	0.1074	0.245	0.1074	0.245	0.1393	0.309
B_2	0.5230	0.976	0.5274	0.992	0.5386	1.023	0.6041	1.140
B_3	0.1264	0.205	0.1308	0.220	0.1180	0.187	0.0689	0.087
GCA_1	0.0706	0.144	0.0760	0.163	0.1260	0.270	0.1361	0.232
GCA_2	-.1077	-.180	-.1186	-.220	-.2186	-.434	-.2371	-.493
GCA_3	-.1316	-.347	-.1426	-.386	-.2426	-.601	-.3972	-.952
GCA_4	0.2489	0.398	0.2544	0.417	0.3044	0.524	0.4241	0.804
GCA_5	0.1265	0.489	0.1320	0.509	0.1820	0.616	0.1746	0.646
SCA_{12}	0.0665	0.172	0.0763	0.208	0.1663	0.400		
SCA_{13}	-.3374	-.628	-.3277	-.592	-.2377	-.400		
SCA_{14}	-.0484	-.128	-.0550	-.152	-.1150	-.280	-.2041	-.410
SCA_{15}	0.0766	0.126	0.0700	0.102	0.0100	-.026	0.0480	0.094
SCA_{23}	0.3995	0.912	0.3600	0.771				
SCA_{24}	0.1528	0.289	0.1627	0.324	0.2527	0.517	0.1920	0.408
SCA_{25}	-.3185	-.706	-.3084	-.670	-.2187	-.478		
SCA_{34}	-.0592	0.164	-.0493	0.129	0.0406	0.064	0.1163	0.246
SCA_{35}	0.3580	0.677	0.3679	0.712	0.4793	0.905		

a) where (numerical examples are for height)

$b_4 = -\sum_1^3 b_i = 0.7697$;

$gca_6 = -\sum_1^5 gca_j = -0.2067$;

$sca_{p6} = -\sum sca_{jk}$ for j or k = p and p = 1,2,3 then $sca_{16} = 0.2428$,

$sca_{26} = -0.3002$, and $sca_{36} = -0.3608$; $sca_{45} = -\sum_1^9 sca_e = -0.2898$,

e = indepentently estimated sca's 1, . . . ,9;

$sca_{46} = sca_{12} + sca_{13} + sca_{15} + sca_{23} + sca_{25} + sca_{35} = 0.2446$; and

$sca_{56} = sca_{12} + sca_{13} + sca_{14} + sca_{23} + sca_{24} + sca_{34} = 0.1737$.

b) where the linear combinations for parameter estimates are identical to the balanced example.

c) where $sca_{p6} = -\sum sca_{jk}$ for j or k = p and p = 1 to 3; $sca_{45} = -\sum_1^8 sca_e$

e = independently estimated SCA's 1,. . .,8;

$sca_{46} = sca_{12} + sca_{13} + sca_{15} + sca_{25} + sca_{35}$; and

$sca_{56} = sca_{12} + sca_{13} + sca_{14} + sca_{24} + sca_{34}$.

d) where $sca_{16} = -sca_{14} - sca_{15}$, $sca_{26} = -sca_{24}$, $sca_{36} = sca_{26}$,

$sca_{46} = sca_{15}$, $sca_{56} = sca_{14} + sca_{24}, + sca_{34}$, and

sca_{25} = the negative of the sum of the four independently estimated sca's.

e) where for all cases linear combinations for block and gca are the same as in the balanced case.

3.2 Missing Plot

To illustrate the problem of a missing plot, the cross, parent two by parent three, was arbitrarily deleted in block one (as if observation y_{123} were missing). This deletion prompts adjustments to factor matrices in order to analyze the new data set. The new vector of observations (y) now has 59 rows. This necessitates deletion of the row of the design matrix (X_s) in blocs 1 which would have been associated with cross 2 x 3. This is the only matrix alteration required for the analysis. Thus, the resultant X_s matrix has 60 — 1 = 59 rows and 18 columns. With 59 means in y and 18 columns in X_s, the degrees of freedom for error is 41.

Comparisons between results of the analysis (*Table* 2) of the full data set and the data set missing observation y_{123} reveal that for this case the estimates of parameters have been relatively unaffected by the imbalance (magnitudes of GCA's changed only slightly and rankings by GCA were unaffected).

3.3 Missing Cross

Another common form of imbalance in diallel data sets, the missing cross, is examined through arbitrary deletion of the 2 x 3 cross from all blocks, i. e. y_{123}, y_{223}, y_{323}, y_{423} are missing in the data vector. This type of imbalance is representative of a particular cross that could not be made and is therefore missing from all blocks. The matrix manipulations required for this analysis are again presented by factor. For appropriate SCA restrictions, the data vector and design matrix should be ordered so that the p^{th} parent has no missing crosses. Since the labeling of a parent as parent p is entirely subjective, any parent with all crosses may be designated as parent p. The previous labelling directions are necessary since we generate the SCA submatrix as horizontal direct products of the columns of the GCA submatrix; and to account for missing crosses, the horizontal direct product for each particular missing parental combinations are not calculated which sets the missing SCA's to zero. If there is a cross missing from those of the p^{th} parent, we cannot account for the missing cross with this technique (SEARLE, 1987, page 479).

For the mean, block, and GCA submatrices, the adjustment for the missing cross dicates deleting the rows in the submatrices which would have corresponded to the y_{i23}

observations. The SCA submatrix must be reformed since a degree of freedom for SCA and hence a column of the submatrix has been lost. The SCA submatrix is reinstituted from the GCA horizontal direct products (remembering that one cross, 2x3, no longer exists and therefore that product $GCA_2 \times GCA_3$ is inappropriate). Dropping the column for SCA_{23} to zero (SEARLE, 1987) so that the remaining SCA's will sum-to-zero. After that, the reformation is according to the established pattern. With one missing cross there are now 56 observations and hence 56 degrees of fredom available. The columns of the X_s matrix are now: one for the mean, three for block, five for GCA, and eight for SCA for a total of 17 columns. The remaining degrees of freedom for error is 39, matching the correct degrees of freedom ((14—1) x (4—1) = 39).

For the missing cross example $\hat{\mu}$ is no longer equivalent to the mean of the plot means since $\hat{\mu} = 2.5386$ and $\Sigma_{ijk}y_{ijk})/N = 2.5715$ where N = 56 (number of plot means). This is the result of GCA effects which are no longer orthogonal to the mean. Check the $X_s'X_s$ matrix or try estimating factors separately and compare to the estimates when all factors are included in X_s.

If formulae for balanced data (BECKER, 1975; FALCONER, 1981; and HALLAUER and MIRANDA, 1981) are applied to unbalanced data (plot-mean basis) estimates of parameters are no longer appropriate because factors in the model are no longer independent (orthogonal). Applying BECKER's formula which uses totals of cross means for a site ($\bar{y}_{.jk}$) to the missing cross example yields: $gca_1 = 0.2992$, $gca_2 = —0.5649$, $gca_3 = —0.5888$, $gca_4 = 0.4665$, $gca_5 = 0.3552$, and $gca_6 = 0.0219$. These answers are very different in magnitude from those in *table 2* for this example and gca_6 also has a different sign. Employing these formulae in the analysis of unbalanced data is analogous to matrix estimation of GCA's without the other factors in the model which is inappropriate.

3.4 Several Missing Crosses

The concluding example (*Table 2*) is a drastically unbalanced data set resulting from the arbitrary deletion of five crosses (1 x 2, 1 x 3, 2 x 3, 3 x 5, and 4 x 5). The matrix manipulation for this example is an extension of the previous one cross deletion example. Rows corresponding to y_{i12}, y_{i13}, y_{i23}, Y_{i35}, and y_{i45} are deleted from the mean, block and GCA submatrices for all blocks. The SCA matrix (now 4 columns = 10 crosses — 5 —1 = 4 degrees of freedom) is again reformed with only the relevant products of the GCA columns. Counting degrees of freedom (columns of the sum-to-zero design matrix), the means has one, block has three, GCA has five, and SCA has four degrees of freedom for a total of 13. Error has (4—1) (10—7) = 27 degrees of fredom. Totaling degrees of freedom for modeled effects and error yields 40 which equals the number of plot means.

In increasingly unbalanced cases (*Table 2*), the spread among the GCA estimates tends to increase with increasing imbalance (loss of information). This is a general feature of OLS analyses and the basis for the feature is that the spread among the GCA estimates is due to both the innate spread due to additive genetics effects as well as the error in estimation of the GCA's. When there is less information, GCA estimates tend to be more widely spread due to the increase in the error variance associated with their estimation. This feature has been noted (WHITE and HODGE, 1989, page 54) as the tendency to pick as parental winners individuals in a breeding program which are the most poorly tested.

Discussion

After developing the OLS analysis and describing the inherent assumptions of the analysis, there are four important factors to consider in the interpretation of sum-to-zero OLS solutions: (1) the lack of uniqueness of the parameter estimates; (2) the weights given to plot means (y_{ijk}) and in turn site means ($\bar{y}_{.jk}$) for crosses in data sets with missing crosses in parameter estimation; (3) the arbitrary nature of using a diallel mean (perforce a narrow genetic base) as the mean about which the GCA's sum-to-zero; and (4) the assumption that the variance-covariance matrix for the observations (V) is $I\sigma^2_e$.

4.1 Uniqueness of Estimates

Sum-to-zero restrictions furnish what would appear to be unique estimates of the individual parameters, e. g. GCA_1, when, in fact, these individual parameters are not estimable (GRAYBILL, 1976; FREUND and LITTELL, 1981; and MILLIKEN and JOHNSON, 1984). The lack of estimability is again analogous to attempting to solve a set of equations in n unknows with t equations where n is greater than t. Therefore, an infinite number of solutions exist for β.

There are quantities in this system of equations that are unique (estimable), i. e. the estimate is invariant regardless of the restriction (sum-to-zero or set-to-zero) or generalized inverse (no restrictions) used (MILLIKEN and JOHNSON, 1984) and the estimable functions include sum-to-zero GCA and SCA estimates since they are linear combinations of the observations; but, these estimable quantities do not estimate the individual parametric GCA's and SCA's of the overparameterized model (*Equation 4*) since there is no unique solution for those parameters.

4.2 Weighting of Plot Means and Cross Means in Estimating Parameters

With at least one measurement tree in each plot and with plot means as the unit of observation, use of the matrix approach produces the same results as the basic formulae. The weight placed on each plot mean in the estimation of a parameter can be determined by calculating $(X_s'X_s)^{-1}X_s'$ which can be viewed as a matrix of weights W so that *equation 3* can be written as b = Wy. The matrix W has these dimensions: the number of rows equals the number of parameters in β_s and the number of columns equals the number of plot means in y. The i^{th} row of the W contains the weights applied to y to estimate the i^{th} parameter in b ($\hat{b}_i$). In the discussion which follows gca_1 is utilized as $\hat{b}_i$.

If there are no missing plots, the cross mean in every block (y_{ijk}) has the same weighting and weights can be combined across blocks to yield the weight on the overall cross mean ($\bar{y}_{\cdot jk}$). It can be shown that for the balanced numerical example gca_1 is calculated by weighting the overall cross means containing parent 1 by 1/6 and weighting all overall cross means not containing parent 1 by —1/12. *Figure 4* (above the diagonal) demonstrates the weightings on the overall cross means for the balanced numerical example as well as the marginal weighting on the GCA parameters. These marginal weightings are obtained by summing along a row and/or column as one would to obtain the marginal totals for a parent (BECKER,

	GCA1	GCA2	GCA3	GCA4	GCA5	GCA6	
GCA1		1/6 .16667	1/6 .16667	1/6 .16667	1/6 .16667	1/6 .16667	5/6
GCA2	.14583 missing		-1/12 -.08333	-1/12 -.08333	-1/12 -.08333	-1/12 -.08333	-1/6
GCA3	.14583 missing	missing missing		-1/12 -.08333	-1/12 -.08333	-1/12 -.08333	-1/6
GCA4	.18056 .22549	-.10417 .01961	-.10417 -.11765		-1/12 -.08333	-1/12 -.08333	-1/6
GCA5	.18056 .31372	-.10417 -.27451	-.10417 missing	-.06944 missing		-1/12 -.08333	-1/6
GCA6	.18056 .29412	-.10417 .08824	-.10417 -.04902	-.06944 -.29412	-.06944 -.20588		-1/6

Figure 4. — Weights on overall cross means ($\bar{y}_{.jk}$) for the three numerical examples for estimation of GCA_1. The weights for the balanced example (above the diagonal) are presented in both fractional and decimal form. The weights for the one-cross missing and the five-crosses missing are presented as the upper number and lower number, respectively, in cells below the diagonal. The marginal weights on GCA parameters (right margin) do not change although cells are missing.

1975). One feature of sum-tozero solutions is that these marginal weightings will be maintained no matter the imbalance due to missing crosses, as will be seen by considering the numerical examples for a missing cross (*Figure 4* below the diagonal, upper number) and five missing crosses (*Figure 4* below the diagonal, lower number). The marginal weights have remained the same as in the balanced case while the weights on the cross means differ among the crosses containing parent 1 and also among the crosses not containing parent 1. In the five missing crosses example, crosses $\bar{y}_{.24}$ and $\bar{y}_{.26}$ even receive a positive weighting where in the prior examples they had negative weighting.

The expected value in all three examples is GCA_{1s} (for sum-to-zero) despite the apparently nonsensical weightings to cross means with missing crosses; however, the evaluation of the estimates in terms of the original model changes with each new combination of missing cells, i. e. $\bar{y}_{.24}$ and $\bar{y}_{.26}$ have a positive weight in the five missing crosses example in GCA_1 estimation. Whether this type of estimation is desirable with missing cell (cross) means has been the subject of some discussion (Speed, Hocking and Hackney, 1978; Freund, 1980; and Milliken and Johnson, 1984). The data analyst should be aware of the manner in which sum-to-zero treats the data with missing cell means and decide whether that particular linear combination of cross means estimating the parameter is one of interest, realizing that the meaning of the estimates in terms of the original model is changing.

4.3 Diallel Mean

The use of the mean for a half-diallel as the mean around which GCA's sum-to-zero is not satisfactory in that the diallel is the mean of a rather narrow genetically based population, and in particular that the comparisons of interest are not usually confined to the specific parents in a specific diallel on a particular site. A checklot can be employed to represent a base population against which comparison of half-or full-sib families can be made to provide for comparison of GCA estimates from other tests (van Buijtenen and Bridgwater, 1986).

Mathematically, when effects are forced to sum-to-zero around their own mean, the absolute value of the GCA's is reflective of their value relative to the mean of the group. Even if the parents involved in the particular diallel were all far superior to the population mean for GCA, GCA's calculated on an OLS basis would show that some of these GCA's were negative. If the GCA's of the diallel parents were in fact all below the population mean, the opposite and equally undesirable result ensues. For disconnected diallels together on a single site, an OLS analysis would yield GCA estimates that sum-to-zero within each diallel since parents are nested within diallels. Unless the comparisons of interest are only in the combination of the parents in a specific diallel on a specific site, the checklot alternative is desirable.

A method for obtaining the desired goal of comparable GCA's from disconnected experiments, disregarding the problem of heteroscedasticity, is to form a function from the data which yields GCA estimates properly located on the number scale. Such a function can be formed (using GCA_1 as an example from gca_{1s}, the diallel mean, and the checklot mean.

From expectations of the scalar linear model (*Equation 1*),

$$GCA_{1s} = ((p-1)/p)GCA_1 - (1/p)\Sigma_{j=2}^{p}GCA_j + (1/p)\Sigma_{k=2}^{p}SCA_{1k} - (2/(p(p-2)))\Sigma_{j=2}^{p-1}\Sigma_{k=3}^{p}SCA_{jk}; \qquad (5)$$

$$E\{\text{diallel mean}\} = \mu + (\Sigma_{i=1}^{b}B_i)/b + (2/p)\Sigma_{j=1}^{p}GCA_j + (2/(p(p-1)))\Sigma_{j=1}^{p-1}\Sigma_{k=2}^{p}SCA_{jk}; \text{ and}$$

$$E\{\text{checklot mean}\} = \mu + (\Sigma_{i=1}^{b}B_i)/b + \tau;$$

The function used to properly locate GCA_{1rel} (the subscript rel denotes the relocated GCA_{1s}) is $gca_{1rel} = gca_{1s} + (1/2)$ (diallel mean — checklot mean). The expectation of gca_{1rel} with negligible SCA is $GCA_{1rel} = GCA_1 - \tau/2$; and since breeding value equals twice GCA, $BV_{1rel} = BV_1 - \tau$. If SCA is non-negligible then the expectation is:

$$GCA_{1rel} = GCA_1 + (1/(p-1))\Sigma_{k=2}^{p}SCA_{1k} - (1/((p-1)(p-2)))\Sigma_{j=2}^{p-1}\Sigma_{k=3}^{p}SCA_{jk} - \tau/2. \qquad (6)$$

In either case the function provides a reasonable manner by which GCA estimates from disconnected diallels are centered at the same location on a number scale and are then comparable.

4.4 Variance and Covariance of Plot Means

The variances of plot means with unequal numbers of trees per plot are by definition unequal, i.e. $\text{Var}(y_{ijk}) = \sigma^2_p + \sigma^2_w/n_{ijk}$ where $\sigma^2 p$ is plot variance, σ^2_w is the within plot variance and n_{ijk} is the number of observations per plot. Also, if blocks were considered random, there would be additional source of variance for plot means due to blocks (as well as a covariance between plot means in the same block) and this could be incorporated into the V matrix with $\text{Var}(y_{ijk}) = \sigma^2_b + \sigma^2_p + \sigma^2_w/n_{ijk}$. Since the variances of the means in the observation vector are not equal and there is a covariance between the means if blocks are being considered random, best linear unbiased estimates (BLUE) would be secured by weighting each mean by it's true associated variance (SEARLE, 1987, page 316). This is the generalized least squares (GLS) approach as:

$$\mathbf{b} = (\mathbf{X_s}'\mathbf{V}^{-1}\mathbf{X_s})^{-1}\mathbf{X_s}'\mathbf{V}^{-1}\mathbf{y} \qquad (7)$$

The GLS approach relaxes the OLS assumptions of equal variance of and no covariance between the observations (plot means) while still treating genetic parameters as fixed effects. The entries along the diagonal of the V matrix are the variances of the plot means ($\text{Var}(y_{ijk})$) in the same order as means in data vector. The off-diagonal elements of V would be either 0 or σ^2_b (the variance due to the random variable block) for elements corresponding to observations in the same block. BLUE requires exact knowledge of V; if estimates of σ^2_p, σ^2_b, and σ^2_w are utilized in the V matrix, estimable functions of β approximate BLUE.

The OLS assumption that SCA and GCA are fixed effects can also be relaxed to allow for covariances due to genetic relatedness. In particular, the information that means are from the same half- or full-sib family could be included in the V matrix. Relaxation of the zero covariance assumption implies that GCA and SCA are random variables. If GCA and SCA are treated as random variables, then the application of best linear prediction (BLP) or best linear unbiased prediction (BLUP) to the problem would be more appropriate (WHITE and HODGE, 1989, page 64). The treatment of the genetic parameters as random variables is consistent with that used in estimating genetic correlations and heritabilities. The V matrix of such an application would include, in addition to the features of the GLS V matrix, the covariance between full-sib or half-sib families added to the off-diagonal elements in V, i, e. if the first and second plot means in the data vector had a covariance due to relationship, then that covariance is inserted twice in the V matrix. The covariance would appear as the second element in the first row and the first element in the second row of V (V is a symmetric matrix). Also the diagonal elements of V would increase by $2\sigma^2_{gca}$ (the variance due to treating GCA as a random variable) + σ^2_{sca} (the variance due to treating SCA as a random variable).

4.5 Comparison of Prediction and Estimation Methodologies

Which methodology (OLS, GLS, BLP, or BLUP) to apply to individual data bases is somewhat a subjective decision. The decision can be based both on the computational or conceptual complexity of the method and the magnitude of the data base with which the analyst is working. To aid in this decision, this discussion highlights the differences in the inherent properties and assumptions of the techniques.

For all practical purposes the answers from the four techniques will never be equal; however, there are two caveats. First, OLS estimates equal GLS estimates if all the cell means are known with the same precision (variance), SEARLE, 1987, page 490). Otherwise, GLS discounts the means that are known with less precision in the calculations and different estimates result. The second caveat is if the amount of data is infinite, i. e. all cross means are known without error, then all four techniques are equivalent (WHITE and HODGE, 1989, pages 104 to 106). In all other cases BLP and BLUP shrink predictions toward the location parameter(s) and produce predictions which are different from OLS or GLS estimates even with balanced data. During calculations GLS, BLP, and BLUP place less weight on observations known with less precision, which is intuitively pleasing.

With OLS and GLS forest geneticists treat GCA's and SCA's as fixed effects for estimation and then as random variables for genetic correlations and heritabilities. BLP and BLUP provide a consistent treatment of GCA's and SCA's as random variables while differing in their assumptions about location parameters (fixed effects). In BLP fixed effects are assumed known without error (although they are usually estimated from the data) while with BLUP fixed effects are estimated using GLS. BLP and BLUP techniques also contain the assumption that the variance-covariance matrix of the observations is known without error (most often variances must be estimated). In many BLUP applications (HENDERSON, 1974), mixed model equations are utilized interactively to estimate fixed affects and to predict random variables from a data set. A BLUP treatment of fixed effects allows any connectedness between experiments to be utilized in the estimation of the fixed effects. This provides an intuitive advantage of BLUP over BLP in experimentation where connectedness among genetic experiments is available or where the data are so unbalanced that treating the fixed effects as known is less desirable than a GLS estimate of the fixed effects.

An ordering of computational complexity and conceptual complexity from least to most complex of the four methods is OLS, GLS, BLP and BLUP. The latter three methods require the estimation of the variance-covariance matrix of the observations either separately (a priori) or iteratively with the fixed effects. Precise estimation of the variance-covariance matrix for observations requires a great number of observations and the precision of GLS, BLP and BLUP estimations or predictions is affected by the error of estimation of the components of V.

Selection of a method can then be based on weighing the computational complexity and size of the available data base against the advantages offered by each method. Thus, if complexity of the computational problem is of paramount concern, the analyst necessarily would choose OLS. With a small data base (one that does not allow reasonable estimates of variances), the analyst would again choose OLS. With a large data base and no qualms with computational complexity, the analyst can choose between BLP and BLUP based on whether there is sufficient connectedness or imbalance among the experiments to make BLUP advantageous.

Conclusions

Methods of solving for GCA and SCA estimates for balanced (plot-mean basis) and unbalanced data have been presented along with the inherent assumptions of the analysis. The use of plot means and the matrix equations will produce sum-to-zero OLS estimates for GCA and SCA for all types of imbalance. Formulae in the literature which yield OLS solutions for balanced data can yield misleading solutions for unbalanced data because of the loss of orthogonality and also weightings on site means for crosses (or totals) are constants.

GCA's and SCA's obtained through sum-to-zero restriction are not truly estimates of parametric population GCA's and SCA's. There are an infinite number of solutions for GCA's and SCAs from the system of equations as a result of the overparameterized linear model. Yet, if the only comparisons of interest are among the specific parents on a particular site, then the estimates calculated by sum-to-zero restrictions are appropriate. Checklots may be used to provide comparability among estimates derived from disconnected sets.

Having discussed the innate mathematical features of OLS analysis, knowledge of these features should help the data analyst decide if OLS is the most desirable technique for the data at hand. It may be desirable to relax OLS assumptions, which are in all likelihood invalid for the variance-covariance matrix of the observations. This could lead to GLS, BLP or BLUP as better alternatives.

References

BECKER, W. A.: Manual of Quantitative Genetics. Washington State Univ.Press, Pullman, WA. 170 pp., (1975). —— BURLEY, J., BURROWS, P. M., ARMITAGE, F. B. and BARNES, R. D.: Progeny test designs for *Pinus patula* in Rhodesia. Silvae Genet. **15**, 166—173 (1966). —— FALCONER, D. S.: Introduction to Quantitative Genetics. Longman & Co., New York, NY. 340 pp., (1981). —— FREUND, R. J.: The case of the missing cell. Amer. Stat. **34**, 94—98 (1980). —— FREUND, R. J. and LITTELL, R. C.: SAS for Linear Models. SAS Institute, Inc., Cary, NC. 231 pp., (1981). —— GILBERT, N. E. G.: Diallel cross in plant breeding. Heredity **12**, 477—498 (1958). —— GRAYBILL, F. A.: Theory and Application of the Linear Model. Duxbury Press, North Scituate, MA. 704 pp., (1976). —— GREENWOOD, M. S., LAMBETH, C. C. and HUNT, J. L.: Accelerated breeding and potential impact upon breeding programs. In: Southern Cooperative Series Bulletin No. 309. Louisiana Ag. Experiment Station, Baton Rouge, LA. pp. 39—41 (1986). —— GRIFFING, B.: Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. **9**, 463—493 (1956). —— HALLAUER, A. R. and MIRANDA, J. B.: Quantitative Genetics in Maize Breeding. Iowa State Univ. Press, Ames, IO. 468 pp., (1981). —— HENDERSON, C. R.: General flexibility of linear model techniques for sire evaluation. J. Dairy Sci. **57**, 963—972 (1974). —— MATZINGER, D. F., SPRAGUE, G. F. and COCKERHAM, C. C.: Diallel crosses of maize in experiments repeated over locations and years. Crop Sci. **51**, 346—350 (1959). —— MILLIKEN, G. A. and JOHNSON, D. E.: Analysis of Messy Data I, Designed Experiments. Lifetime Learning Pub., Belmont, CA. 473 pp., (1984). —— SAS INSTITUTE, INC.: SAS Interactive Matrix Language Guide for Personal Computers. SAS Institute, Inc., Cary, NC. 429 pp., (1985). —— SCHNEIDER, D. M.: Linear Algebra, a Concrete Introduction. Maxmillan Pub Co., New York. 506 pp., (1987). —— SEARLE, S. R.: Linear Models for Unbalanced Data. John Wiley and Sons, New York, NY. 536 pp., (1987). —— SINGH, M. and SINGH, R. K.: A comparison of different methods of half- diallel analysis. Theor. Appl. Genet. **67**, 323—326 (1984). —— SNYDER, E. B. and NAMKOONG, G.: Inheritance in a diallel crossing experiment with longleaf pine. In: USDA For. Serv. Res. Pap. SO-140. South. For. Exp. Stn., New Orleans, LA. 31pp., (1978). —— SPEED, M. M., HOCKING, R. R. and Hackney, O. P.: Methods of analysis of linear models with unbalanced data. J. Amer. Stat. Assoc. **73**, 105—112 (1978). —— SPRAGUE, G. F. and TATUM, L. A.: General vs. specific combining ability in single crosses of corn. J. Amer. Soc. Agron. **34**, 923—932 (1942). —— SQUILLACE, A. E.: Comparison of some alternative second-generation breeding plans for slash pine. In: South. For. Tree Improve. Conf. June 12—13, 1973 Baton Rouge, LA, pp. 2—13 (1973). —— VAN BUIJTNEN, J. P. and BRIDGWATER, F.: Mating and genetic test designs. In: Advanced Generation Breeding of Forest Trees. Southern Coop. Series Bull. 309. Louisiana Ag. Exp. Stn., Baton Rouge, LA. pp. 5—10 (1986). —— WEIR, R. J. and GODDARD, R. E: Advanced generation operational breeding programs for loblolly and slash pine. In: Southern Coop. Series Bull. 309. Lousiana Agric. Exp. Stn., Baton Rouge, LA. pp. 21—26 (1986). —— WEIR, R. J. and ZOBEL, B. J.: Managing genetic resources for the future a plan for the N. C. State Industry Cooperative Tree Improvement Program. In: Proc. 13th South For. Tree Improve. Conf. June 10 to 11, Raleigh, NC. pp. 73—82 (1975). —— WHITE, T. L. and HODGE, G. R.: Predicting Breeding Values with Applications in Forest Tree Improvement. Kluwer Academic Pub., Dordrecht, The Netherlands. 367 pp., (1989). —— WILCOX, M. D., SHELBOURNE, C. J. A. and FIRTH, A.: General and specific combining ability in eight selected clones of radiata pine. N. Z. J. For. Sci. **5**, 219—225 (1975). YATES, F.: The analysis of multiple classifications with unequal numbers in the different classes. J. Amer. Stat. Assoc. **29**, 51—66 (1934).

Segregation and Linkage of Allozymes in Seed Tissues of the Hybrid Greek Fir Abies borisii regis Mattfeld

By B. FADY[1]) and M. T. CONKLE[2])

(Received 30th December 1991)

Summary

Seed tissues (haploïd megagametophyte and diploïd embryo tissue) of *Abies borisii regis* were used in starch gel electrophoresis to study inheritance and linkage of isozyme variants. The 10 enzyme systems studied are coded by a minimum of 15 isozyme loci. All loci code allozymes in both megagametophyte and embryo tissues. Mendelian segregation ratios were found for all enzyme systems except ACO and 6-PGD where distortion was observed. Segregation distortion could also exist in other enzyme systems (LAP, PGI2, MNR1, MNR2). Evidence of total linkage is provided for one pair of loci (GR/MNR1) that has never been tested before in conifers and tight linkage for another pair of loci (MNR1/PGI1).

Key words: *Abies borisii regis*, allozymes, inheritance, linkage, electrophoresis, seed tissues.

[1]) INRA, Unité Expérimentale d'amélioration des Arbres Forestiers Méditerranéens, Domaine du Ruscas, 4935, Route du Dom, F-83237 Bormes les Mimosas Cedex

[2]) Institute of Forest Genetics, USDA Forest Service, 1960 Addison St, P. O. Box 245, Berkeley, CA 94701, USA