To the Graduate Council:

I am submitting herewith a dissertation written by Donald McLean Panter, Jr. entitled "Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree Doctor of Philosophy, with a major in Plant and Soil Science.

Fred L. Allen

Fred L. Allen, Major Professor

We have read this dissertation
and recommend it acceptance:

Robert A. McLean

Dennis West

Henry H. Fortney

Accepted for the Council:

Carolyn Winkel

Associate Vice Chancellor
and Dean of the Graduate School

Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Donald McLean Panter, Jr.

August 1991

DEDICATION

This dissertation is dedicated to my wife, Jennifer. No man could be as blessed as I am to be loved and supported by such a wonderful woman. I love you.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the staff and faculty of the Department of Plant and Soil Science at the University of Tennessee, Knoxville, for making my experience there one that I will cherish. In particular, I would like to thank the past and current members of the Soybean Breeding team: Dr. Harbans Bhardwaj, Dr. Wayne Whitehead, Dr. Amma Nimako-Bruce, Debbie Ellis, Mark Sellmann, Alan Lowman, and Daqi Qian. I have enjoyed working with each and every one of you.

To the members of my doctoral committee, Dr. Bob McLean, Dr. Dennis West, Dr. Henry Fribourg, and Dr. Vern Reich, I wish to express my thanks for their patience during the preparation of this manuscript.

I reserve this final note of thanks for the person who took a chance on me as an undergraduate student, and then guided me through my graduate career: Dr. Fred Allen. There is little that I can write here that would adequately express the admiration and gratitude I feel in my heart. Thanks for everything, Doc.

ABSTRACT

Best linear unbiased predictions (BLUP) from mixed linear models have been used to predict breeding values of dairy sires for milk yield based on information from their dams and daughters. One of the major advantages of BLUP is that predictions of individuals can be made when information on the individual *per se* is unavailable, but information from its relatives is available. Since BLUP methodology can utilize information from individuals *per se* and their relatives to predict the value of individuals, there is potential for its use in two important areas of plant breeding: i) predicting breeding values of parents from the performance of their relatives, and ii) ranking new genotypes when observed data for them are limited (*e.g.* early stages of performance testing). The objectives of this dissertation were to i) compare the efficiencies of BLUP and MPV in determining seed yield performances of future soybean (*Glycine max* (L.) Merr.) crosses from historical information about parents, ii) determine the effects of progeny and grand-progeny yield performance information on breeding values of their parents, iii) compare seed yield predictions from BLUP to a traditional approach of estimation, best linear unbiased estimations (BLUE), for ranking new genotypes from a limited number of yield tests, and iv) develop a computing strategy for utilizing BLUP in plant breeding applications. The F_4-F_6 bulks and $F_{5:6}$ lines from 24 crosses and four parents were evaluated in replicated yield trials in 11 environments to establish their relative seed yield performances. A

summary of the results was: i) predictions of the 24 crosses from BLUP using only historical parental data were better indicators of performance than estimates from MPV, ii) BLUP breeding values of parents were more precise using small amounts of progeny information than when using large amounts of grand-progeny information, iii) BLUP was superior to BLUE for ranking new genotypes evaluated in a limited number of performance trials, and iv) computer software was developed using SAS/IML™ for computing BLUP values in plant breeding applications. The BLUP methodology should be considered a superior alternative to traditional approaches to genotypic performance estimation.

TABLE OF CONTENTS

PART III: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

II. Influence of Performance of Relatives on Breeding Values of Parents

PART IV: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

III. Selection of Superior Progeny from a Limited Number of Yield Trials

PART V: A Computing Strategy for Utilizing Mixed Linear Models and BLUP in Plant Breeding Applications

PART VI: Overall Summary

LIST OF TABLES

TABLE PAGE

PART II: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

I. Choosing Parents

PART III: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

II. Influence of Performance of Relatives on Breeding Values of Parents

PART IV: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

III. Selection of Superior Progeny from a Limited Number of Yield Trials

LIST OF FIGURES

FIGURE PAGE

PART II: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

I. Choosing Parents

PART III: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

II. Influence of Performance of Relatives on Breeding Values of Parents

PART IV: Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

III. Selection of Superior Progeny from a Limited Number of Yield Trials

PART V: A Computing Strategy for Utilizing Mixed Linear Models and BLUP in Plant Breeding Applications

PART I

General Introduction and Overview

CHAPTER 1

Introduction and Literature Review

Plant breeders historically have faced the problem of parental selection in order to obtain hybrid populations with high expected mean performance and significant genetic variation. Identification of cross combinations which meet the above criteria result in higher probabilities of selecting progenies with sufficient genetic superiority to warrant potential variety release. Typically, predicting the mean of cross combinations is made by calculating midparent values (MPV), or the mean of parental means, based on observed records of potential parents. In order to have the most confidence in these MPV as predictors of future progeny, it is important to obtain the best unbiased estimate of a parent's mean as possible.

Achieving the best estimate of a single trait mean often is accomplished by employing some form of an additive linear model. With such a model, effects which contribute to observed values can be "adjusted out" and an "unbiased" estimate of the genotypic effect can be calculated. A classical method of obtaining the adjusted estimate is by considering all effects fixed and combining data from field tests across locations and/or years. This approach results in ordinary or weighted least squares

solutions (White *et al.*, 1986). The effects due to locations, years, and blocks within location/year are "adjusted out" and the best linear unbiased estimate (BLUE) of the genotypic effect is isolated. However, in cases where only unbalanced data (individuals are not equally represented across years, locations, and blocks) are available, the fixed effects approach can lead to imprecise genotypic estimates resulting in a tendency to select parents which are included in fewer tests (Henderson, 1973; White *et al.*, 1986).

Several methods of multivariate analysis (Grafius, 1965; Pederson, 1981; Whitehouse, 1968; Riggs, 1973) have been developed to simplify selection of potential parents. These selection indices require that estimates of trait means be as unbiased as possible. Obtaining unbiased estimates can be accomplished as described above under the restriction of relatively balanced data. These methods of parental selection share two assumptions: 1) progeny means equal MPV, and 2) none of the potential parents are related. The latter assumption means that breeding values of the individual are determined by observations of the individual itself, and that the addition of information on relatives of the individual has no effect on its predicted breeding value. Bhatt (1970; 1973) addressed the question of covariance between parents using an inverse matrix of error mean squares and crossproducts to calculate a generalized distance between any two potential parents. The emphasis of Bhatt's work was to identify potential parents which were most genotypically diverse. Although this method provides a means to identify diversity between two individuals, it cannot

supplement predictions of breeding values with information on relatives of the individual.

Henderson (1975a; 1977) described the use of a mixed linear model to calculate the best linear unbiased predictions (BLUP) of breeding values of potential parents based on observed records and known (or estimated) variance/covariance structure among fixed and among random effects. Using mixed linear models, genotypic effects can be considered random and other effects, such as years and locations, can be considered fixed. Henderson's methods utilize the genetic relationship among individuals as the variance/covariance among genetic effects and assume that the only cause of correlation between observations on different individuals is additive genetic variance (Henderson, 1975b). By utilizing the genetic variance/covariance among individuals, related individuals contribute to one another's predicted values. He demonstrated later (Henderson, 1977) how a narrow sense estimate of heritability of the trait of interest also could be incorporated into the model. In cases of balanced data and no correlation (*i.e.* relationships) among genotypes, the mixed model solutions (BLUP's) and least squares solutions (BLUE's) lead to the same relative ranking of genotypes (White *et al.*, 1986).

The use of genetic relationships among individuals provides a significant advantage to plant breeders since observations of related individuals contribute to the predictions of each other. First, when few or no observations of an individual are available, information from relatives of the individual can contribute to its predicted breeding value. Second, the magnitude of the contribution is determined by the extent

of the relationship between two individuals. For example, information on a sister-line of an individual would contribute more to its predicted breeding value than information on a distant relative.

Another property of BLUP which is advantageous to plant breeders is the shrinkage of predictions of individuals for which there is little information (Hill and Rosenberger, 1985; Stroup, 1989). This shrinkage tends to "adjust" predictions from the observed mean of the individual and toward the overall mean when there are small numbers of observations of a particular individual, and "adjust" predictions closer to the observed individual mean as the number of observations increases. The error variance of BLUP predictions is lower than for conventional methods of breeding value predictions (Hill and Rosenberger, 1985; White *et al.*, 1986). White *et al.* (1986) demonstrated that in cases where there are few observations, least squares predictions are least reliable and the errors associated with these predictions are greatest. Conversely, when BLUP's are made on few observations, the variances among the predictions are small, and the errors associated with these predictions are least. Since plant breeders often want to predict breeding values for many individuals which have not been tested extensively for performance, BLUP should have a significant advantage over other methods of predicting breeding values.

Henderson's (1975a) mixed linear model can be represented in matrix form by the following equation:

$$y = \mu + X\beta + ZU + \epsilon \quad [1]$$

where

y is an n x 1 vector of measured responses (*e.g.* yield),

μ is the overall mean of the measured response (*e.g.* mean yield)

X is an n x p design matrix of fixed effects (*e.g.* years, locations),

ß is the p x 1 vector of unknown fixed effects to be estimated (*e.g.* estimated effect for Location 1, Location 2, *etc.*)

Z is an n x q design matrix of random effects (*e.g.* genotypes),

U is the q x 1 vector of unknown random effects, with a mean of zero, to be predicted (*e.g.* predicted effect of Genotype 1, Genotype 2, *etc.*)

ϵ is an n x 1 error vector with a null mean.

For the above model,

n is the number of observations,

p is the number of fixed effects, and

q is the number of random effects.

This model assumes that the variances and covariances of the vectors U and ϵ are given by the matrices G and R respectively. In many genetic applications, G is equal to the matrix of genetic relationships among individuals. The matrix R is equal to the variance/covariance among errors. In cases where there is no correlation among errors, R is a diagonal matrix and easily invertible. In the following matrix equations, an apostrophe (') indicates the transpose of a matrix, and a superscript minus (-) indicates the generalized inverse of a matrix. Henderson demonstrated that

estimates of U and ß (u and b, respectively) can be obtained from any solution of the following equations:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} * \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}. \quad [2]$$

Solving for the unknown vector yields the equations:

$$\begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix} * \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+G^{-1} \end{bmatrix}^{-}. \quad [3]$$

For genetic applications when the heritability (h^2) of the trait has been estimated from an unselected, noninbred population, the equations become:

$$\begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix} * \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\frac{1-h^2}{h^2} \end{bmatrix}^{-}. \quad [4]$$

where

A equals a matrix of genetic relationships among individuals being evaluated (Henderson, 1977).

There appear to be several deterrents in the use of BLUP. First, there is a general lack of researchers with experience in BLUP techniques. However, recent literature has been published pertaining to the utility of Henderson's BLUP

methodology which should provide the necessary background for more researchers to utilize mixed linear models. McLean (1989) provided a comparison between fixed and mixed models methodology where he demonstrated the differences in techniques and principles. Sanders (1989) described how to choose and compute the parameter estimates for both fixed and mixed models analyses. Hill and Rosenberger (1985) showed that BLUP was superior to other methods of combining unbalanced data for estimating mean yields of alfalfa (*Medicago sativa* L.) genotypes from a series of evaluations at one location over a period of eight years. Stroup (1989) gave a description of the "shrinkage estimator" of BLUP using batting averages of professional baseball players from three dates in 1985. He demonstrated that BLUP from early season batting averages were better predictors of whole-season batting averages than BLUE from late-season batting averages. Bridges (1989) reported that BLUP was superior to BLUE in estimating the mean yields of four cucumber (*Cucumis sativus* L.) cultivars evaluated at two locations using a latin square design.

Henderson (1977) provided information of the utility of BLUP for predicting mean performance when genotypes are considered random effects and the variance/covariance structure among random effects is equal to the genetic relationships among individuals. In dairy cattle, sires have a major influence on milk production, but the genetic potential of a sire cannot be measured directly. He demonstrated how the breeding values for milk yield of sires could be predicted from the performance of their daughters using BLUP. Allen and Bhardwaj (1987) used BLUP to evaluate the actual *versus* predicted response of soybean (*Glycine max* (L.)

Merr.) ancestral lines and cultivars to simulated acid rain. Allen *et al.* (1991) later demonstrated, conceptually, how Henderson's mixed models methodology could be applied to plant breeding and used to predict mean seed yields of future soybean populations. They used simulated data to illustrate how the genetic relationships among hypothetical soybean cultivars could be used to provide feasible estimates of progeny from crosses among the cultivars. They also showed how predictions from BLUP should theoretically be superior to predictions made from MPV.

The second limitation in the use of BLUP is the lack of computer hardware available to make the analyses efficient. In equations [3] and [4] above, solving for the vector [b / u] requires the inversion of the right hand matrix containing four sub-matrices. This matrix has the dimensions p+q by p+q, where p equals the number of fixed effects and q equals the number of random effects. In cases where many effects and their interactions are being evaluated, inversion of this matrix becomes impossible on many microcomputers and expensive or impossible on many mainframe computers. Currently, the only feasible way to address this problem is to keep the number of fixed and random effects limited to analyses which will not extend beyond the computer resources available or write computer programs which take advantage of any specific matrix structure (*e.g.* block diagonality) that may exist.

Another area where computer hardware is a limitation is in the creation and subsequent inversion of the A matrix in genetic applications. In many plant breeding programs where the matrix of genetic relationships among hundreds of individuals must be created and inverted in order to make predictions, computer hardware

powerful enough to perform these tasks is not available. Henderson (1975c) developed a computing algorithm which replaces "brute force" in the creation and inversion of the A matrix. His FORTRAN program accepts data as a list of progeny and their parents, and creates the elements of A and/or A^{-1} as the data are read, as opposed to the physical inversion of a large matrix. This program can be an invaluable tool for plant breeders in the computation of genetic relationships. However, this algorithm does not account for self-fertilization in plants and subsequent changes in inbreeding following several generations of selfing.

Another factor which has restricted the use of BLUP is the lack of computer software to perform mixed models analyses. Only one program is currently available which can perform mixed models analyses with relative ease (Blouin *et al.*, 1989). However, this program, called GLMM, cannot incorporate a genetic variance/covariance matrix into the mixed models equations. Currently, there is no "canned" computer software available for genetic applications where the genetic relationships among the individuals to be predicted are to be incorporated in the mixed models equations. Until such programs become available, it will be necessary to write explicit computer instructions to be able to perform mixed models analysis. Fortunately, programs can be written to perform mixed models analyses in matrix language modules such as SAS/IML™ (SAS, 1985), APL, and Gauss (McLean, 1989).

Breeding soybeans for yield may serve as a model system for the use of BLUP. First, Henderson's mixed models analyses account only for additive genetic variance. Yield in soybeans has been shown to be primarily under additive genetic

control (Gates, 1960; Brim, 1961; Croissant, 1971). Second, current soybean germplasm has been shown to trace back to relatively few ancestral lines (Allen and Bhardwaj, 1987) making information about genetic relationships among potential parents extremely important. The use of the genetic relationship matrix should effectively increase the amount of information available on an individual when observations of its close relatives are available. Finally, as with any breeding program, prediction of breeding values of potential parents is complicated by having unequal numbers of observations at different locations in different years. Since BLUP analysis adjusts predictions and the errors associated with them based on the number of observations available, the bias normally connected with evaluating genotypes with few observations should be reduced.

CHAPTER 2

Objectives

There were several objectives in this study:

1) to develop computer software in SAS/IML™ capable of performing mixed models analyses for genetic applications where the genetic relationships among individuals of interest could be incorporated into the analyses as the variance/covariance among random genetic effects,

2) to develop computer software in SAS/IML™ to calculate the genetic relationships among individuals from a data set containing a list of pedigrees using the computer algorithm provided by Henderson (1975c) while adding the ability to account for subsequent generations of inbreeding,

3) to compare predictions from BLUP to estimates from MPV in evaluating the mean seed yield of future soybean crosses from historical parental data,

4) to determine the effects of performance information from progeny and close relatives on the predicted seed yield breeding values of their parents using BLUP, and

5) to determine the relative efficiencies of BLUE and two methods of BLUP in estimating or predicting mean seed yield of late-generation soybean crosses from evaluations conducted in only one or two environments.

CHAPTER 3

Overview

This dissertation is divided into four different manuscripts, each addressing one or more of the objectives stated above. The objective of the first manuscript, entitled "Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Yield in Soybeans: I. Choosing Parents", is to determine if BLUP is superior to MPV for choosing parents for soybean seed yield crosses when only historical seed yield performance information on the parents is available to estimate their breeding values. Historical data compiled on the parents is used to estimate their breeding values and to predict the mean seed yields of crosses between them using both BLUP and MPV. Predictions from BLUP and MPV are compared to the means of actual crosses evaluated for seed yield in field experiments for three years to compare the relative efficiencies of each method for predicting seed yield cross means. This paper illustrates how a plant breeder might utilize BLUP to reduce significantly the number of crosses needed to identify superior progeny.

The second manuscript, entitled "Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Yield in Soybeans: II. Influence of Performance of

Relatives on Breeding Values of Parents", refines the concepts established in the first manuscript by demonstrating how a breeder might use BLUP to predict breeding values of parents based on information from their progeny. Yield performance information on the direct progeny of two different parents is used to demonstrate the additive effects of very close relatives on the breeding values of each parent; and yield performance information on the grand-progeny of another parent is used to show the additive effects of more distantly related relatives on the breeding value of a parent. The paper also illustrates how much information on progeny of a parent is needed to obtain an accurate prediction for the breeding value of a parent.

After crosses have been made and lines are established from those crosses, a breeder must evaluate the performance of the new lines in order to select superior material. Cost constraints usually dictate the number of different environments that a breeder can use to evaluate new lines. The third manuscript, entitled "Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Yield in Soybeans: III. Selection of Superior Progeny from a Limited Number of Yield Trials", demonstrates how BLUP could be utilized by a breeder to make accurate predictions of the ranks for seed yield of new lines from fewer evaluation trials. It also shows how historical data on the parents of new lines could be incorporated into the prediction methodology in order to enhance the predicted values of new lines.

The final manuscript, entitled "A Computing Strategy for Utilizing Mixed Linear Models and Best Linear Unbiased Predictions in Plant Breeding Applications", is intended to be published as a Tennessee Agricultural Experiment Station bulletin.

The purposes of this paper are to describe and demonstrate the computing strategy that can be employed for utilizing BLUP, and to present computer programs written to calculate BLUP in genetic applications. It is hoped that this paper will help provide breeders with a vital component of BLUP which has been missing to date: software capable of performing mixed models analyses for genetic applications.

Literature Cited

Allen, F.L. and H.L. Bhardwaj. 1987. Genetic relationships and selected pedigree diagrams of North American soybean cultivars. Tennessee Agric. Exp. Stn. Bull. 652.

Allen, F.L., D.M. Panter, W.L. Sanders, and R.A. McLean. 1991. Best linear unbiased prediction as a method to enhance breeding for yield in soybeans. Crop Sci. (in review).

Bhatt, G.M. 1970. Multivariate analysis approach to selection of parents for hybridization aiming at yield improvement in self-pollinated crops. Aust. J. Agric. Res. 21:1-7.

Bhatt, G.M. 1973. Comparison of various methods of selecting parents for hybridization in common bread wheat (*Triticum aestivum* L.). Aust. J. Agric. Res. 24:457-464.

Blouin, D.C., A.M. Saxton, and K.L. Koonce. 1989. A completely randomized design with missing cells and repeated measures in time. p. 80-103. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Bridges, Jr., W.C. 1989. Analysis of a plant breeding experiment with heterogeneous variances using mixed model equations. p. 145-154. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Brim, C.A. and C.C. Cockerham. 1961. Inheritance of quantitative characters in soybeans. Crop Sci. 1:187-190.

Croissant, G.L. and J.H. Torrie. 1971. Evidence of nonadditive effects and linkage in two hybrid populations of soybeans. Crop Sci. 11:675-677.

Gates, C.E., C.R. Weber, and T.W. Horner. 1960. A linkage study of quantitative characters in a soybean cross. Agron. J. 52:45-49.

Grafius, J.E. 1965. A geometry of plant breeding. Michigan State Univ. Agric. Exp. Stn. Res. Bull. No. 7, East Lansing, MI.

Henderson, C.R. 1973. Sire evaluation and genetic trends. p. 10-41. *In* Animal Breeding and Genetics Symposium in Honor of J.L. Lush, ASAS and ADSA, Champaign, IL.

Henderson, C.R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-477.

Henderson, C.R. 1975b. Use of all relatives in intraherd prediction of breeding values and producing abilities. J. Dairy Sci. 58:1910-1916.

Henderson, C.R. 1975c. Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci. 58:1727-1730.

Henderson, C.R. 1977. Prediction of future records. p. 615-638. *In* E. Pollack *et al.* (ed.). Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames, IA.

Hill, R.R. and Rosenberger, J.L. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470.

McLean, R.A. 1989. An introduction to general linear models. p. 9-38. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

McLean, R.A., W.L. Sanders, and Walter Stroup. 1990. A unified approach to mixed linear models. J. Am. Stat. Assoc. 45:54-64.

Pederson, D.G. 1981. A least-squares method for choosing the best relative proportions when intercrossing cultivars. Euphytica 30:153-160.

Riggs, T.J. 1973. The use of canonical analysis for selection within a population of spring barley. Ann. Appl. Biol. 74:249.

Sanders, W.L. 1989. Choice of models. p. 49-56. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

SAS Institute Inc. 1985a. SAS® Users's Guide: Statistics, Version 5 Edition. Cary, NC. SAS Institute Inc. 956 pp.

SAS Institute Inc. 1985b. SAS/IML™ Users's Guide for Personal Computers, Version 6 Edition. Cary, NC. SAS Institute Inc. 243 pp.

Stroup, W.W. 1989. Why mixed models? p. 1-8. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

White, T.L., G.R. Hodge, and M.A. Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. p. 99-122. *In* Workshop of the Genetics and Breeding of Southern Forest Trees RIEG, June 25-26, 1986, Univ. of Florida, Gainsville, FL.

Whitehouse, R.N.H. 1968. An application of canonical analysis to plant breeding. p. 61. *In* Proc. 5th Congress Eucarpia, Milano.

PART II

Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

I. Choosing Parents

Abstract

Selecting parental cross combinations is one of the most important factors in a plant breeding program. Choosing parents to cross typically is accomplished by estimating parental breeding values from historical data and calculating the midparent value (MPV) for potential crosses. In cases where either limited or no data exist for parents of interest, precise predictions are difficult to obtain. A preferred method of parental selection should be able to identify superior cross combinations from historical parental data even in cases where data are not available on some parents. One such method, best linear unbiased predictions (BLUP), has been utilized to determine paired matings in dairy cattle. The objectives of this study were to compare the efficiencies of two methods of parental selection, MPV and BLUP, for identifying superior soybean (*Glycine max* (L.) Merr.) cross combinations in cases when i) equal and unequal amounts of yield performance data on all potential parents were available, and ii) unequal amounts of yield performance data were available on some parents and no data were available on other parents. The F_4-F_6 bulks (11 environments) and the $F_{5:6}$ lines (5 environments) from 24 soybean crosses were evaluated in replicated trials to estimate the mean seed yield performance of each cross. Historical records on the yield performance of the parents of each cross were compiled from yield trials conducted in Tennessee from 1982-1990 and were used to predict the performance of the 24 crosses. Numbers of records on the parents were

restricted to provide simulated situations of balanced and unbalanced parent data availability. The performance of each cross was predicted using MPV and BLUP for each situation. Standard error of the differences ($s_{\bar{d}}$) and rank correlations between the actual predicted performances of the 24 crosses were calculated to determine the relative efficiencies of MPV and BLUP. In every case, predictions from BLUP provided higher rank correlations, lower $s_{\bar{d}}$, and identified higher percentages of the superior crosses than MPV. When there were no data on some parents, BLUP provided more precise predictions of the 24 crosses than MPV when all data on all parents were available. These results indicate that BLUP should be considered a superior alternative to MPV for selecting parental cross combinations.

CHAPTER 1

Introduction and Literature Review

One of the most important, yet difficult, decisions breeders face is choosing parents to combine which have the highest probability of producing superior offspring. Ideally, a breeder would like to employ methodology that used historical performance data on potential parents and could identify superior cross combinations from among the potential parents. This type of system could conserve resources that would have been required for evaluating inferior crosses. In cases when there is either limited or no performance data on some of the parents of interest, accurate predictions of the performance of future progeny are very difficult to obtain using classical approaches to parental selection.

Several methods of multivariate analysis (Grafius, 1965; Pederson, 1981; Whitehouse, 1968; Riggs, 1973) have been developed to simplify selection of potential parents. These methods of parental selection share the assumptions that progeny means equal midparent values (MPV) and that none of the potential parents are related. The latter assumption means that breeding values of the individual are determined by observations of the individual itself, and that the addition of

information on relatives of the individual has no effect on its predicted breeding value. Bhatt (1970, 1973) used an inverse matrix of error mean squares and crossproducts to calculate a generalized distance between any two potential parents. The emphasis of Bhatt's work was to identify potential parents which were most genotypically diverse. Although this method provides a means to identify diversity between two individuals, it cannot supplement predictions of breeding values with information on relatives of the individual.

Henderson (1975a, 1977) described the use of a mixed linear model to calculate the best linear unbiased predictions (BLUP) of breeding values of potential parents based on observed records and known (or estimated) variance/covariance structure among random effects. Henderson's methods utilize the genetic relationships among individuals as the variance/covariance among random genetic effects and assume that the only cause of correlation between observations on different individuals is additive genetic variance (Henderson, 1975b). By utilizing the genetic variance/covariance among individuals, related individuals contribute to one another's predicted values. He demonstrated later (1977) how a narrow sense estimate of heritability of the trait of interest also could be incorporated into the model. In cases of balanced data and no correlation among genotypes, the mixed model solutions (BLUP) and least squares solutions, best linear unbiased estimates (BLUE) lead to the same relative ranking of genotypes (White *et al.*, 1986).

The use of genetic relationships among individuals provides a significant advantage to plant breeders since observations of related individuals contribute to the

predictions of each other. First, when few or no observations of an individual are available, information from relatives of the individual can contribute to its predicted breeding value. Conversely, the traditional approach of calculating MPV cannot predict the values of potential crosses made with parents for which there is no data. Second, the magnitude of the contribution is determined by the extent of the relationship between two individuals. For example, information on a sister-line of an individual would contribute more to its predicted breeding value than information on a distant relative.

Another property of BLUP which is advantageous to plant breeders is the shrinkage of predictions of individuals for which there is little information (Hill and Rosenberger, 1985; Stroup, 1989). This shrinkage tends to "adjust" predictions back toward the overall mean when there are small numbers of observations on a particular individual, and "adjust" predictions closer to the observed individual mean as the number of observations increases. The error variance of BLUP is lower than for conventional methods of breeding value predictions (Hill and Rosenberger, 1985; White *et al.*, 1986). White *et al.* (1986) demonstrated that in cases where there are few observations, least squares predictions are least reliable and the errors associated with these predictions are greatest. Conversely, when BLUP are made on few observations, the variances among the predictions are small, and the errors associated with these predictions are least. Since plant breeders often want to predict breeding values for many individuals which have not been tested extensively

for performance, BLUP should have a significant advantage over other methods of predicting breeding values.

Until recently, the lack of published details on the effectiveness of BLUP may have been a deterrent in the use of BLUP in plant breeding. However, papers by Stroup (1989), McLean(1989), Sanders (1989), Bridges (1989), and Blouin *et al.* (1989) describing the potential uses for BLUP in many areas of agricultural research have helped provide the necessary background information researchers need to be able to apply this methodology. Hill and Rosenberger (1985) demonstrated the superiority of BLUP over other methods of combining unbalanced data for estimating mean yields of alfalfa (*Medicago sativa* L.) genotypes from a series of evaluation trials conducted at one location over a period of eight years.

Henderson (1973) reported the first use of BLUP for purely genetic applications where the genetic relationships among individuals were utilized in the mixed models equations. He used BLUP successfully to predict the breeding values for milk yield of dairy sires based on the performance of their daughters. Allen and Bhardwaj (1987) utilized BLUP to evaluate the predicted *versus* actual yield response of soybean ancestral lines and cultivars to simulated acid rain. Allen *et al.* (1991) used simulated data to demonstrate, conceptually, how BLUP could be used to predict parental breeding values of soybean strains in cases of balanced and unbalanced parental data, and how BLUP compared to traditional MPV.

Another factor which has restricted the use of BLUP is the lack of computer software to perform the analysis. Blouin *et al.* (1989) developed a computer

program, GLMM, which virtually eliminates the need to write computer code to perform most types of mixed models analyses. Currently, GLMM is restricted to analyses where no covariances exist among random effects, and therefore genotypes must be considered to be unrelated. Until "canned" programs become available, it will be necessary to write explicit computer instructions to be able to perform mixed models analysis for genetic applications. Fortunately, programs can be written to perform mixed models analyses in matrix language modules such as SAS/IML™ (SAS, 1985b), APL, and Gauss (McLean, 1990).

Breeding soybeans for yield may serve as a model system for the use of BLUP. First, Henderson's mixed models analyses account only for additive genetic variance. Yield in soybeans has been shown to be primarily under additive genetic control (Gates, 1960; Brim, 1961; Croissant, 1971). Second, current soybean germplasm has been shown to trace back to relatively few ancestral lines (Allen and Bhardwaj, 1987) making information about genetic relationships among potential parents extremely important. The use of the genetic relationship matrix should effectively increase the amount of information available on an individual when observations of its close relatives are available. Finally, predicting breeding values of potential parents is complicated by having unequal numbers of observations at different locations in different years. Since BLUP adjusts predictions and the errors associated with them based on the number of observations available, the bias normally connected with evaluating breeding values of potential parents with few observations should be reduced.

The objectives of this study were to determine if i) BLUP is superior to MPV for predicting a single quantitative trait, seed yield, of future soybean crosses from historical parental data when equal amounts of yield performance data on all potential parents were available, and ii) if BLUP provided precise predictions of seed yield performance of future soybean crosses in "real-world" situations when unequal amounts of yield performance data were available on some parents and no data were available on other parents.

CHAPTER 2

Materials and Methods

Twenty-four unselected F_4 soybean crosses representing diverse genetic backgrounds were chosen as entries for this study (Table 1[1]). The 24 crosses were not predetermined for this study, but were a sample of a larger group of crosses processed yearly in the Univ. of Tennessee Soybean Breeding Program (UTSB). The single criterion for selection was that there be enough seed available for a cross to be entered into a replicated yield trial. The crosses were derived from hand-pollinations made in 1986. The F_1 plants were grown at the Knoxville Exp. Stn. (KES) in the summer of 1987. The F_2 seed were harvested and advanced through modified single seed descent in the F_2 and F_3 generations at a winter nursery facility in Belize, Central America.

[1] Tables and figures are presented in the Appendix.

Field Performance Trials

The 24 crosses were evaluated for seed yield as F_4 bulks in two-row plots, 6 m long, with 90-cm row spacing, in a randomized incomplete block design with four replications in 1988 at KES. Both rows of each plot were trimmed to 4.9 m and harvested for seed yield. In 1989, each F_5 bulk cross was grown in yield tests at five Tennessee locations: KES, Highland Rim Exp. Stn. (HRES), Ames Plantation (AMES), Middle Tennessee Exp. Stn. (MTES), and Milan Exp. Stn. (MES). Each yield test consisted of four-row plots, 6 m long, 90-cm row spacing, and was arranged in a randomized incomplete block design with four replications. At harvest, the two center rows of each plot were trimmed to 4.9 m and harvested for seed yield.

At each of four locations in 1989, KES, MES, AMES, and HRES, 12 individual F_5 plants were selected at random from the border rows of each cross (*i.e.* 48 $F_{5:6}$ lines total per cross). In 1990, 40 of the selected $F_{5:6}$ lines from each cross were tested for seed yield; eight lines being assigned at random to each of five locations: KES, MES, AMES, HRES, and MTES. Plots were four rows, 3 m long, 90-cm row spacing, and arranged in a randomized incomplete block design with four replications. Each plot consisted of two $F_{5:6}$ lines from one of the crosses in the two center rows and the two border rows consisted of F_6 bulk seed from the same cross. At harvest, one bulk row (border) and both $F_{5:6}$ lines were harvested individually from each plot and seed yield and moisture were recorded. All yields

were adjusted to 130 g H_20 kg^{-1}. The purpose of these yield trials was to establish rankings of the 24 crosses from the best possible estimates of mean seed yield for each cross based on tests of bulks and lines across a broad range of environments.

Historical yield data on the parents of each cross were compiled from the UTSB and the USDA Uniform Soybean Tests - Southern Region conducted in Tennessee (only) from 1982 to 1990. These data consisted of the mean seed yield of a parent from each trial in which the parent occurred. Since some parents had been tested for yield more extensively than others, there was a wide range in the number of records on each parent (Table 2). Three parents, Hutcheson, TN 4-86, and TN 5-85 made up 34% of the total number of historical parent observations; and three parents, Hutcheson, TN 4-86, and TN85-55 were parents of at least six of the 24 crosses being evaluated. Many of the parents were highly related. For example, 24% and 28% of the parents were derived from crosses with Essex and D72-8927, respectively.

Statistical Analyses

Each year/location combination was considered to be a separate environment. Overall least squares mean yields were calculated for both bulks (11 environments) and lines (5 environments) for each cross. All effects were considered fixed. Using rank correlation analysis, it was determined that bulk and line data could be

combined to provide the best estimate of overall mean performance of each cross. These estimates were used as the target values for comparing MPV and BLUP.

Based on pedigree information presented by Allen and Bhardwaj (1987), genetic relationships among parents and progeny were determined using a SAS/IML™ program (D.M. Panter and R.A. McLean, 1990, unpublished) which is comparable to Henderson's (1975c) FORTRAN program.

Using the historical records on the parents, the mean seed yield of each cross was predicted using two different methods. For both prediction methods it was assumed that i) a progeny of a cross between two parents received half of its genes from each parent, and ii) all parents were homozygous and homogeneous.

The first method of prediction was to calculate MPV of crosses from parental mean performance. Considering all effects fixed, least squares mean performance of parents was estimated across all historical data and the MPV of each cross was equal to the mean of parental least squares means. This method represented a "classical" technique of predicting progeny performance.

The second method was to use mixed models analysis to make BLUP for each cross using the same historical parental data as were used for MPV. The additive linear mixed model was:

$$y = \mu + X\beta + ZU + \epsilon ,$$

where

y is a vector of measured yield,

μ is the overall mean yield,

X is a design matrix of fixed environments,

ß is the vector of unknown effects due to environments to be estimated,

Z is a design matrix of random parents and progeny,

U is the vector of unknown effects due to parents and progeny, and

ϵ is an error vector with a null mean.

Since BLUP required an estimate of heritability for the given trait, variance components were estimated from 1990 $F_{5:6}$ line data using the REML option in SAS® PROC VARCOMP (SAS, 1985a) and heritability for yield (0.24) was calculated based on these estimates. BLUP values were calculated using a SAS/IML™ program (D.M. Panter, 1990, unpublished).

One of the major advantages of BLUP is its ability to predict precisely the values for a group of individuals when the number of observations on each individual is not equal (Henderson, 1975b; Hill and Rosenberger, 1985). However, in situations where equal numbers of observations on each individual are available, it was not clear whether BLUP retained its advantage over classical methods of prediction. For BLUP to be considered superior to MPV, its predictions of the mean yield of future crosses must be consistently better in cases of balanced data in order to compensate for the added costs accrued in performing the BLUP analysis.

To compare the effectiveness of BLUP *versus* MPV when equal amounts of data were available for each parent, eight of the nine parents for which there were at least nine historical records were selected (Table 2). One potential parent,

TN85-157, which had more than nine historical records, was not involved in any crosses with the other eight parents and was not used in the analysis. Of the total 56 potential diallel crosses from the eight parents, seven had been evaluated in field trials (1988-1990). The means of these seven crosses were predicted using both BLUP and MPV based on historical data sets where 1, 3, 6, and 9 random records were available for each of the eight selected parents. Limiting the predictive data sets to the eight selected parents insured that records from the other 17 parents would not enhance the BLUP of the simulated crosses; there would have been no effect on the MPV of the simulated crosses since MPV utilizes only information on the parents *per se* of a cross. By keeping the number of records on each parent equal in each set of predictions, it was assured that any shrinkage of predictions made by BLUP would be due to the covariances among parents, and not due to unequal numbers of observations on some parents.

Standard errors of the difference ($s_{\bar{d}}$) between the actual and the predicted means of the seven crosses were calculated to determine the accuracy of BLUP *versus* MPV when equal information was available on all parents. The $s_{\bar{d}}$ in each case provided a measure of the variance in prediction errors associated with each method of prediction. The method with the lowest $s_{\bar{d}}$ would be considered to provide more precise, and therefore, superior predictions.

In practice, it would be highly unlikely that a breeder would have access to equal amounts of information for all parents that potentially could be mated. A more likely situation would be where many records were available on some parents,

and few or no records were available on other parents. By randomly deleting some of the parents in the historical parent data set, all of these situations could be simulated. In order to determine the effectiveness of BLUP in predicting and ranking cross means when little or no data were available on some parents, 100% (*i.e.* control), 75%, 50%, and 25% of the 25 parents were selected at random from the historical parental data and BLUP of cross means were calculated based on the assumed data availability. For example, when data were available on 100% of the parents, the predictive parental data set contained all records on all 25 parents; when it was assumed that data were available on 75% of the parents, the predictive set contained all records on 19 random parents, *etc*. Since there must be at least one record for each parent to calculate MPV, a comparison between MPV and BLUP could only be made at the 100% level of assumed parent availability in these analyses. The $s_{\bar{d}}$ and rank correlations (r_s) between actual and predicted cross means were calculated for each set of predictions. Counts were made to ascertain how many of the top i) 12 crosses and ii) 6 crosses overall (Table 1) would have been identified by selecting the top 12 and 6 crosses from the BLUP or MPV predictions. The rationale was to determine which method of prediction could rank most effectively the superior cross combinations so that the best parental choices could have been made *a priori*.

Using the 1990 data on lines from the 24 crosses, environment and block within environment effects were estimated. The yield of each line from each cross was adjusted by subtracting the corresponding environment and block within

environment estimates from the observed yield to provide the least biased estimate of the performance of the individual line. Using this scheme, lines from different environments and blocks within environments could be compared to one another for yield performance. Based on the adjusted yield for all lines, transgressively segregating lines were identified as those lines which had adjusted yields greater than one and two standard deviations ($\hat{\sigma}$=540 kg ha^{-1}) from the overall adjusted mean yield ($\hat{\mu}$=2040 kg ha^{-1}). Counts were made to determine what percentage of the transgressively segregating lines would have been identified if crosses were made based on predictions from BLUP *versus* MPV.

Based on the yield data from the lines, genetic variance was estimated for each cross and was correlated with the genetic covariance of the parents of the cross. The purpose of this analysis was to determine if an increase in genetic variation could be expected by crossing more genetically diverse parents.

CHAPTER 3

Results and Discussion

Bulk *versus* Line Means

Rank correlation between least squares mean seed yields across environments for bulk (F_4-F_6) and $F_{5:6}$ line data was found to be positive (r_s=0.77) and highly significant ($P \leq 0.01$). Therefore, data for both bulks and lines were combined to provide the best estimate attainable for each cross from the total 11 different environments. The overall least squares mean seed yield (Table 1) was considered the target value for each method of prediction in later analyses. The top 12 crosses (overall ranks 1-12 based on measured yield, Table 1) and top 6 crosses (overall ranks 1-6, Table 1) were considered to be target crosses that should be made if a breeder wanted to make only 50% or 25% of the total of 24 crosses.

MPV *versus* BLUP for Predicting Future Cross Means with Equal Information on Each Parent

The predicted mean yields of seven simulated crosses when only 1, 3, 6, and 9 random records for each of the eight selected parents were available are presented in Table 3. The highest yielding cross based on actual data, CROSS-23, was predicted to be the highest yielding cross by both methods in every case. In three out of four cases (1, 6, and 9 random records for each parent), BLUP predicted CROSS-23 at least 109 kg ha^{-1} lower than MPV even though the predictive data were the same in each case. These observations reflect one of the important properties of BLUP in genetic applications: the contributions of relatives to the predictions of breeding values of individuals.

Examination of the genetic variance/covariance matrix among the eight parents reveals that the parents of CROSS-23, Hutcheson and TN85-55, share at least some of their genes with all other parents (Table 4). The breeding value of Hutcheson is affected most by observations on Hutcheson *per se*, but is also affected to a certain degree by two close relatives, TN 5-85 and TN85-117 (genetic covariances with Hutcheson of 0.33 and 0.29, respectively), which yield notably less than Hutcheson (2782 and 2510 *versus* 3065 kg ha^{-1}, respectively, Table 4). These relatives, among others, tended to decrease the predicted breeding value of Hutcheson. The same principle holds true for TN85-55, the other parent of CROSS-23. One of its close relatives, TN 4-86 (genetic covariance of 0.28) yields

much less than TN85-55 (2299 *versus* 2791 kg ha^{-1}, respectively, Table 4) and therefore has the effect of decreasing the breeding value of TN85-55. These complex interactions have the overall effect of decreasing the predicted value of CROSS-23.

This principle was also demonstrated for CROSS-11, a cross between TN85-117 and TN 4-86. In every case, the BLUP of CROSS-11 was higher than the MPV. Since Hutcheson was a high-yielding relative of TN85-117, the predicted breeding value of TN85-117 was adjusted upward from its observed value. Likewise, since TN85-55 was a high-yielding relative of TN 4-86, the breeding value of TN 4-86 was adjusted upward as well. The overall effect was the increase in the predicted mean of CROSS-11. These observations coincide with those of Allen *et al.* (1991).

The $s_{\bar{d}}$ between the actual (from field trials) and predicted means ranged from 76 to 68 kg ha^{-1} for MPV and from 63 to 52 kg ha^{-1} for BLUP (Table 3). The largest $s_{\bar{d}}$ for both methods was when only one record was available for each parent. In every case, the $s_{\bar{d}}$ for BLUP was less than for MPV, demonstrating that the variance in errors of prediction with BLUP were always less than those of MPV. These observations are in agreement with those of Hill and Rosenberger (1985), White *et al.* (1986), and Stroup (1989). One very important observation was that, in this analysis, the $s_{\bar{d}}$ for BLUP when only one record was available for each parent was actually less than the best case scenario for MPV when all nine records for each parent were available (63 *versus* 72 kg ha^{-1}, respectively, Table 3). This observation

indicates that predictions from BLUP with little predictive data may be more precise than those from MPV with significantly more data. Stroup (1989) demonstrated similar concepts of BLUP in predicting full season batting averages from early season performance of professional baseball players.

From an applied plant breeding standpoint, each of these observations has very important implications. First, when a breeder evaluates an individual for a quantitative trait predominantly controlled by additive genetic effects, it could be considered that common "portions" (*i.e.* genes in common) of many related individuals are being evaluated at the same time. The magnitude of those "portions" is determined by the degree of genetic relationships existing between the individual being evaluated and its relatives. If the breeder utilizes BLUP as described here, the known genetic covariances between pairs of individuals can be exploited to evaluate genetic rather than genotypic additive effects as MPV does. The overall effect of BLUP is to increase the effective number of observations for each individual by utilizing the records on the individual *per se* as well as those of its relatives.

The second important implication of these analyses is that a breeder can expect that predictions from BLUP will be more precise than from MPV when equal amounts of information are available for potential parents. When future crosses are predicted, they are generally ranked in order of descending priority based on the objectives of the breeding program. The predicted yield levels *per se* are of limited importance due to the relatively large environmental effects, but the ranks are critical in order to choose the best cross combinations to make. As such, prediction

errors are only minimally important, but variances among those errors determine the precision of the method of prediction. Results from these analyses clearly illustrate the superiority of BLUP over MPV in making more precise predictions of future crosses.

BLUP Rankings of Crosses with Partial Parental Data

The $s_{\bar{d}}$ between actual and predicted cross means remained relatively constant (30 to 33 kg ha^{-1}) and rank correlations decreased slightly (r_s=0.63 to 0.52) as the percentage of parents used to predict the BLUP of cross mean seed yield decreased from 100% to 25% of the total of 25 parents (Table 5). The only direct comparison between BLUP and MPV that could be made was at 100% (or all 25 parents) assumed parent availability. At 100%, the $s_{\bar{d}}$ was lower (30 *versus* 38 kg ha^{-1}) and the r_s was higher (0.63 *versus* 0.42) for BLUP than for MPV, respectively. The superiority of BLUP can be attributed to better predictions of the breeding values of the 25 parents. For example, only two records were available for TN85-116, a line derived from a cross between Essex and D72-8927 (Table 2). Using MPV, only those two records would have been used to predict the breeding value of TN85-116 (2014 kg ha^{-1}, Table 4). When BLUP was utilized, data on all relatives contained in the data set also affected the predicted breeding value of TN85-116. Specifically, 13 and 4 records were available for its sister lines, TN85-117 and TN85-13,

respectively (Table 2). Their observations enhanced the prediction for TN85-116 by effectively increasing its number of observations. In the same manner, any parents who shared common ancestry with TN85-116 had an influence on its breeding value, and *vice versa*. This example is just one of the many complex relationships that were exploited when future cross means were predicted using BLUP.

Predictions from BLUP when 75, 50, and 25% of the 25 parents were available were compared to estimates from MPV when 100% of the parents were available (best case scenario). These comparisons showed that the $s_{\bar{d}}$ was always lower and the r_s was always higher for BLUP than for MPV. At the lowest level of parent data availability (25% or 7 parents), BLUP still provided more precise predictions and better rank correlations than MPV. One reason for the precision of BLUP under these conditions is that BLUP uses the genetic relationships among individuals, as described earlier, to increase the effective number of observations of related individuals. Because many of the parents were highly related (Table 2), data from parents included in the predictive set adjusted predictions of the missing parents and their corresponding progeny in correlation with their genetic relationships with one another. The more closely related the parents were, the more effect an included parent had on the prediction of a missing parent or a progeny of a missing parent. Since there were no actual data on missing parents, their predicted values were adjusted toward the overall mean, thereby making these predictions fairly conservative.

The ability of BLUP to identify superior crosses was not diminished under conditions of unavailable data on some parents. With the least percentage of parent data available (25%), 58% of the top 12 crosses would have been identified by selecting the top 12 crosses based on BLUP, while 67% would have been identified using MPV with all (100%) parent data available (Fig. 1). Likewise, at least 50% of the top 6 crosses were identified for each level of parent data availability using BLUP, compared to 33% when MPV was used with all parental data. Notably, the best predictions of the top 6 crosses occurred when only 25% of the parents were available. This anomaly occurred because Hutcheson, one of the top yielding parents overall and a parent of each of the top 6 crosses, was randomly selected for the predictive set.

These observations supplement the rank correlation analyses described earlier, but they also provide information about the sensitivity of BLUP for identifying superior crosses. Choosing crosses *a priori*, based on BLUP would result in more superior crosses than when using traditional MPV. The BLUP methodology was applied to only 24 cross combinations in this study and a higher percentage of the superior crosses was consistently identified than when MPV was used. In practice, BLUP could be used to simulate hundreds of paired matings among a set of potential parents. The combinations with the greatest potential could be predicted in advance, thereby decreasing the number of empirical combinations made. Furthermore, the predictions could lead to identification of superior combinations which otherwise would have gone unnoticed.

Analysis of Lines within Crosses

All crosses produced transgressively segregating lines (Table 6). Among the top 12 crosses overall, there were 64% of the transgressive segregates greater than $\hat{\mu}+1\hat{\sigma}$ and 62% greater than $\hat{\mu}+2\hat{\sigma}$. These percentages revealed that the majority of superior lines was derived from superior crosses. Therefore, a prediction method which identifies superior crosses should also maximize the potential for selecting superior transgressive segregates from the crosses.

When BLUP was used to predict the top 12 crosses based on all available parental data, 67% of the transgressive segregates greater than $\hat{\mu}+1\hat{\sigma}$ and 73% greater than $\hat{\mu}+2\hat{\sigma}$ would have been identified *versus* 59 and 65% for MPV, respectively. These results indicate that BLUP was superior to MPV in identifying crosses with the most potential for producing superior lines.

The correlation of genetic variance of a cross and the genetic relationship of parents of the cross was negative ($r=-0.47$) and significant ($P \leq 0.05$). There was no association between yield level and genetic variance ($r=0.01$ NS). These observations agree with plant breeding theory that combining genetically diverse parents produces greater genetic variation among the progeny.

CHAPTER 4

Conclusions

These results demonstrate that BLUP is superior to MPV in identifying superior cross combinations under a wide range of circumstances. The variances of prediction errors were lower and the rank correlations were higher for BLUP than for MPV in every analysis. One clear advantage of BLUP over MPV is that a cross combination can be predicted with BLUP when no data are available on the parent(s) but are available on relatives (ancestors and/or offspring from other matings). When data on some parents were unavailable, BLUP provided predictions of future soybean cross mean yields as precisely or better than under the best case scenario of MPV. This study shows that using BLUP can increase the efficiency of identifying parental cross combinations with the highest probabilities of producing superior lines.

There is potential for the application of this methodology in several areas of plant breeding, but particularly for parental selection. One possible scheme for using BLUP could be:

i) predict the mean performance of progeny from all combinations of a group of potential parents,

ii) select the combinations with the highest predicted mean,

iii) within the selected group, identify those combinations for which the parents are most genetically diverse (from the genetic relationship matrix), and

iv) make crosses.

Literature Cited

Allen, F.L. and H.L. Bhardwaj. 1987. Genetic relationships and selected pedigree diagrams of North American soybean cultivars. Tennessee Agric. Exp. Stn. Bull. 652.

Allen, F.L., D.M. Panter, W.L. Sanders, and R.A. McLean. 1991. Best linear unbiased prediction as a method to enhance breeding for yield in soybeans. Crop Sci. (in review).

Bhatt, G.M. 1970. Multivariate analysis approach to selection of parents for hybridization aiming at yield improvement in self-pollinated crops. Aust. J. Agric. Res. 21:1-7.

Bhatt, G.M. 1973. Comparison of various methods of selecting parents for hybridization in common bread wheat (*Triticum aestivum* L.). Aust. J. Agric. Res. 24:457-464.

Blouin, D.C., A.M. Saxton, and K.L. Koonce. 1989. A completely randomized design with missing cells and repeated measures in time. p. 80-103. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Bridges, Jr., W.C. 1989. Analysis of a plant breeding experiment with heterogeneous variances using mixed model equations. p. 145-154. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Brim, C.A. and C.C. Cockerham. 1961. Inheritance of quantitative characters in soybeans. Crop Sci. 1:187-190.

Croissant, G.L. and J.H. Torrie. 1971. Evidence of nonadditive effects and linkage in two hybrid populations of soybeans. Crop Sci. 11:675-677.

Gates, C.E., C.R. Weber, and T.W. Horner. 1960. A linkage study of quantitative characters in a soybean cross. Agron. J. 52:45-49.

Grafius, J.E. 1965. A geometry of plant breeding. Michigan State Univ. Agric. Exp. Stn. Res. Bull. No. 7, East Lansing, MI.

Henderson, C.R. 1973. Sire evaluation and genetic trends. p. 10-41. *In* Animal Breeding and Genetics Symposium in Honor of J.L. Lush, ASAS and ADSA, Champaign, IL.

Henderson, C.R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-477.

Henderson, C.R. 1975b. Use of all relatives in intraherd prediction of breeding values and producing abilities. J. Dairy Sci. 58:1910-1916.

Henderson, C.R. 1975c. Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci. 58:1727-1730.

Henderson, C.R. 1977. Prediction of future records. p. 615-638. *In* E. Pollack *et al.* (ed.). Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames, IA.

Hill, R.R. and Rosenberger, J.L. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470.

McLean, R.A. 1989. An introduction to general linear models. p. 9-38. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

McLean, R.A., W.L. Sanders, and Walter Stroup. 1990. A unified approach to mixed linear models. J. Am. Stat. Assoc. 45:54-64.

Pederson, D.G. 1981. A least-squares method for choosing the best relative proportions when intercrossing cultivars. Euphytica 30:153-160.

Riggs, T.J. 1973. The use of canonical analysis for selection within a population of spring barley. Ann. Appl. Biol. 74:249.

Sanders, W.L. 1989. Choice of models. p. 49-56. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

SAS Institute Inc. 1985a. SAS® Users's Guide: Statistics, Version 5 Edition. Cary, NC. SAS Institute Inc. 956 pp.

SAS Institute Inc. 1985b. SAS/IML™ Users's Guide for Personal Computers, Version 6 Edition. Cary, NC. SAS Institute Inc. 243 pp.

Stroup, W.W. 1989. Why mixed models? p. 1-8. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

White, T.L., G.R. Hodge, and M.A. Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. p. 99-122. *In* Workshop of the Genetics and Breeding of Southern Forest Trees RIEG, June 25-26, 1986, Univ. of Florida, Gainsville, FL.

Whitehouse, R.N.H. 1968. An application of canonical analysis to plant breeding. p. 61. *In* Proc. 5th Congress Eucarpia, Milano.

Appendix

Table 1. Mean yields and ranks of 24 soybean crosses based on tests of F_4, F_5, and F_6 bulks (1988-1990), $F_{5:6}$ lines (1990), and bulks and lines combined.

		Yield			Rank		
Soybean Cross	I.D.	Bulks†	Lines‡	Overall§	Bulks	Lines	Overall
		------ kg ha^{-1} ------					
Hutcheson/TN82-221	CROSS-22	2645	1882	2589	2	2	1
Hutcheson/TN85-55	CROSS-23	2657	1838	2584	1	3	2
TN85-116/Hutcheson	CROSS-10	2489	1983	2514	4	1	3
D83-3318/Hutcheson	CROSS-24	2548	1828	2506	3	4	4
TN 4-86/Hutcheson	CROSS-04	2372	1806	2378	5	5	5
TN85-3/Hutcheson	CROSS-18	2347	1643	2310	6	8	6
TN84-3/TN85-157	CROSS-06	2310	1651	2287	7	7	7
TN85-102/TN84-21	CROSS-09	2286	1663	2274	8	6	8
TN85-55/TN84-3	CROSS-21	2282	1633	2262	9	9	9
TN85-172/TN85-157	CROSS-16	2254	1574	2224	10	15	10
TN85-162/TN85-172	CROSS-15	2194	1628	2200	13	10	11
TN85-55/TN 5-85	CROSS-19	2237	1502	2190	11	18	12
Stafford/TN 4-86	CROSS-03	2127	1604	2146	17	11	13
TN85-136/TN85-55	CROSS-14	2150	1537	2141	15	17	14
TN85-121/TN85-116	CROSS-12	2115	1596	2135	18	12	15
TN84-7/TN85-55	CROSS-07	2212	1377	2133	12	22	16
TN85-117/TN 4-86	CROSS-11	2104	1579	2123	19	14	17
TN85-55/TN83-26	CROSS-20	2164	1439	2120	14	19	18
TN85-13/TN 4-86	CROSS-13	2077	1564	2099	20	16	19
TN82-221/TN84-7	CROSS-05	2145	1350	2079	16	23	20
TN85-10/TN 4-86	CROSS-08	2071	1432	2054	21	20	21
D83-3349/TN 4-86	CROSS-01	2005	1401	1999	22	21	22
TN85-22/TN 4-86	CROSS-17	1846	1585	1947	24	13	23
D83-3349/V81-141	CROSS-02	1966	1311	1944	23	24	24

† Least squares means of three replications at one location in 1988 and five locations in 1989 and 1990 (11 environments).
‡ Least squares means of 40 lines (eight lines at each of five locations (environments)) in 1990.
§ Least squares means of bulks at 11 environments and lines at five environments (1988-1990).

Table 2. Mean yield and background information of the 25 parents used in the 24 crosses in this study.

Parent	Pedigree	Historical Yield Records	Parent in Crosses	Yield†	SE
		----- no. -----		-- kg ha^{-1} --	
D83-3318	Peking/Centennial// Forrest///Bedford	4	1	2078	244
D83-3349	Peking/Centennial// Forrest///Bedford	6	2	2142	211
Hutcheson	V68-1034/Essex	32	6	2750	95
Stafford	V66-318/V68-2331	25	1	2547	105
TN 4-86	Beford/Crawford	24	7	2423	107
TN 5-85	D68-127/Essex	31	1	2728	92
TN82-221	J74-45/Ts72-824	8	2	2582	214
TN83-26	J74-40/K1017	13	1	2694	138
TN84-21	Bedford/RA 401	6	1	2630	195
TN84-3	J74-45/V76-482	6	2	2317	195
TN84-7	J74-45/V76-482	6	2	2154	196
TN85-10	D77-18/Fayette	13	1	2502	137
TN85-102	Jeff/Nathan	8	1	2565	170
TN85-116	Essex/D72-8927	2	2	2014	327
TN85-117	Essex/D72-8927	13	1	2523	137
TN85-121	D72-8927/Epps	7	1	2366	182
TN85-13	Essex/D72-8927	4	1	2515	235
TN85-136	AS 5618/D72-8927	4	1	2633	235
TN85-157	D72-8927/TN80-83	16	2	2634	129
TN85-162	Fayette/TN80-57	4	1	3003	235
TN85-172	TN80-70/D72-8927	4	1	2019	237
TN85-22	AS 5474/D72-8927	4	1	2230	235
TN85-3	D77-18/PI 416.772	4	1	1743	235
TN85-55	TN77-46/Fayette	9	6	2688	161
V81-141	Essex/V71-793	2	1	2057	342

† Least squares means and and standard errors (SE) based on historical yield records.

Table 3. Actual† and predicted (MPV and BLUP) mean yields and standard errors of the difference ($s_{\bar{d}}$ = actual *versus* predicted) of seven soybean crosses with varying numbers of records (1, 3, 6, and 9) available for each of eight parents selected from this study.

Cross		Predicted Value with Varying No. of Records on Each Parent							
		1 Record		3 Records		6 Records		9 Records	
I.D.	Actual	MPV	BLUP	MPV	BLUP	MPV	BLUP	MPV	BLUP
CROSS-03‡	2146	2587	2534	2654	2756	2249	2413	2223	2387
CROSS-04	2378	2553	2568	2822	2880	2736	2630	2682	2765
CROSS-08	2054	2688	2581	2766	2845	2535	2560	2413	2501
CROSS-11	2123	2184	2499	2610	2899	2454	2506	2404	2509
CROSS-19	2190	2386	2588	3058	3008	2854	2712	2787	2734
CROSS-20	2120	2587	2582	3035	3000	2675	2641	2754	2683
CROSS-23	2584	2856	2627	3114	3131	2947	2739	2928	2819
$s_{\bar{d}}$		76	63	73	56	68	57	72	52

† Least squares means of F_4 bulks at one locations in 1988; F_5 and F_6 bulks at five locations in 1989-1990; and $F_{5:6}$ lines at five locations in 1990.
‡ See Table 1 for pedigrees.

Table 4. Seed yield and genetic variance/covariance coefficients for pairwise combinations of the eight selected parents for which at least nine historical records of yield performance were available.

Parent	Seed Yield†	Parent							
		TN 5-85	TN 4-86	Hutcheson	Stafford	TN85-55	TN85-10	TN85-117	TN83-26
		variance/covariance							
TN5-85	2782	1.11‡							
TN4-86	2299	0.19§	1.01						
Hutcheson	3065	0.33	0.06	1.02					
Stafford	2146	0.02	0.11	0.07	1.00				
TN85-55	2791	0.11	0.28	0.04	0.08	1.00			
TN85-10	2527	0.15	0.30	0.06	0.08	0.41	1.01		
TN85-117	2510	0.37	0.09	0.29	0.02	0.06	0.10	1.04	
TN83-26	2717	0.19	0.38	0.06	0.11	0.23	0.25	0.09	1.01

† Least squares means based on nine random records chosen from the complete parental data set.
‡ Coefficients along the diagonals of the matrix are variances (Allen and Bhardwaj, 1987).
§ Coefficients on the off-diagonals of the matrix are covariances (Allen and Bhardwaj, 1987).

Table 5. Actual (overall†) and predicted (MPV and BLUP) mean yields, ranks, rank correlations (r_s) and standard errors of the difference ($s_{\bar{d}}$=actual *versus* predicted) of 24 soybean crosses simulating data available on 100 (control), 75, 50, and 25% of the parent lines used in this study.

			Varying Percentages of the 25 Parents in the Predictive Data Set									
	Overall		MPV		BLUP							
Cross I.D.	Yield	Rank	100	Rank	100	Rank	75	Rank	50	Rank	25	Rank
	kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}	
CROSS-22‡	2589	1	2666	4	2627	4	2689	5	2653	5	2768	4
CROSS-23	2584	2	2719	1	2702	1	2780	1	2694	1	2792	1
CROSS-10	2514	3	2382	17	2559	7	2647	7	2609	10	2767	5
CROSS-24	2506	4	2414	16	2504	12	2571	15	2645	7	2750	6
CROSS-04	2378	5	2587	7	2593	6	2674	6	2685	2	2788	2
CROSS-18	2310	6	2246	22	2417	19	2500	21	2636	8	2745	7
CROSS-06	2287	7	2476	11	2481	17	2566	16	2534	18	2613	22
CROSS-09	2274	8	2598	6	2545	9	2618	10	2560	17	2658	19
CROSS-21	2262	9	2503	9	2523	10	2607	12	2576	13	2649	20
CROSS-16	2224	10	2327	19	2442	18	2611	11	2458	24	2658	18
CROSS-15	2200	11	2511	8	2557	8	2728	2	2576	14	2782	3
CROSS-19	2190	12	2708	2	2682	2	2643	8	2669	3	2697	14
CROSS-03	2146	13	2485	10	2511	11	2582	13	2617	9	2723	8
CROSS-14	2141	14	2661	5	2617	5	2705	4	2577	12	2708	9
CROSS-12	2135	15	2190	23	2394	22	2479	22	2489	22	2666	17
CROSS-07	2133	16	2421	15	2491	13	2625	9	2525	19	2589	23

Table 5 (continued)

			Varying Percentages of the 25 Parents in the Predictive Data Set									
	Overall		MPV		BLUP							
Cross I.D.	Yield	Rank	100	Rank	100	Rank	75	Rank	50	Rank	25	Rank
	kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}		kg ha^{-1}	
CROSS-11	2123	17	2473	12	2485	15	2559	19	2572	16	2700	13
CROSS-20	2120	18	2691	3	2663	3	2709	3	2666	4	2705	10
CROSS-13	2099	19	2469	13	2483	16	2565	17	2572	15	2700	12
CROSS-05	2079	20	2368	18	2416	20	2534	20	2484	23	2564	24
CROSS-08	2054	21	2463	14	2487	14	2575	14	2651	6	2669	16
CROSS-01	1999	22	2283	21	2370	23	2436	23	2608	11	2683	15
CROSS-17	1947	23	2327	20	2410	21	2560	18	2498	21	2704	11
CROSS-02	1944	24	2099	24	2349	24	2393	24	2502	20	2643	21
$s_{\bar{d}}$			38		30		32		33		32	
r_s			0.42*		0.63**		0.53**		0.45*		0.52**	

*,** Signficant at P $\leq$ 0.05 and 0.01, respectively

† Least squares means of F_4 bulks at one locations in 1988; F_5 and F_6 bulks at five locations in 1989-1990; and $F_{5:6}$ lines at five locations in 1990.

‡ See Table 1 for pedigrees.

Table 6. Overall ranks, genetic standard deviations ($\hat{\sigma}_g$), genetic covariances between parents, and numbers of transgressively segregating lines from 24 crosses.

Cross			Parental	Trans. Segregates¶	
I.D.	Rank‡	$\hat{\sigma}_g$	Covariance§	$\hat{\mu}+1\hat{\sigma}$	$\hat{\mu}+2\hat{\sigma}$
		kg ha^{-1}		--- no. ---	
CROSS-22†	1	478	0.07	7	2
CROSS-23†	2	464	0.04	8	4
CROSS-10†	3	399	0.29	9	2
CROSS-24†	4	506	0.11	7	2
CROSS-04†	5	620	0.06	7	0
CROSS-18	6	460	0.05	3	1
CROSS-06	7	302	0.26	3	0
CROSS-09†	8	310	0.26	3	2
CROSS-21†	9	295	0.13	4	2
CROSS-16	10	294	0.48	3	0
CROSS-15†	11	508	0.09	6	1
CROSS-19†	12	502	0.11	4	0
CROSS-03†	13	602	0.11	4	1
CROSS-14†	14	638	0.03	5	3
CROSS-12	15	348	0.37	4	1
CROSS-07	16	226	0.13	2	1
CROSS-11	17	576	0.09	5	2
CROSS-20†	18	482	0.23	3	0
CROSS-13	19	305	0.09	4	0
CROSS-05	20	411	0.37	2	0
CROSS-08	21	396	0.30	2	0
CROSS-01	22	485	0.48	1	1
CROSS-17	23	523	0.04	2	1
CROSS-02	24	428	0.13	2	0

† Crosses predicted by BLUP as the top yielding 12 crosses.

‡ Rank based on least squares mean yield bulks at 11 environments and lines at five environments (1988-1990).

§ Genetic relationship between the two parents of the cross.

¶ Number of lines with adjusted means greater than one ($\hat{\sigma}$=540 kg ha^{-1}) and two standard deviations from the overall adjusted mean yield ($\hat{\mu}$=2040 kg ha^{-1}).

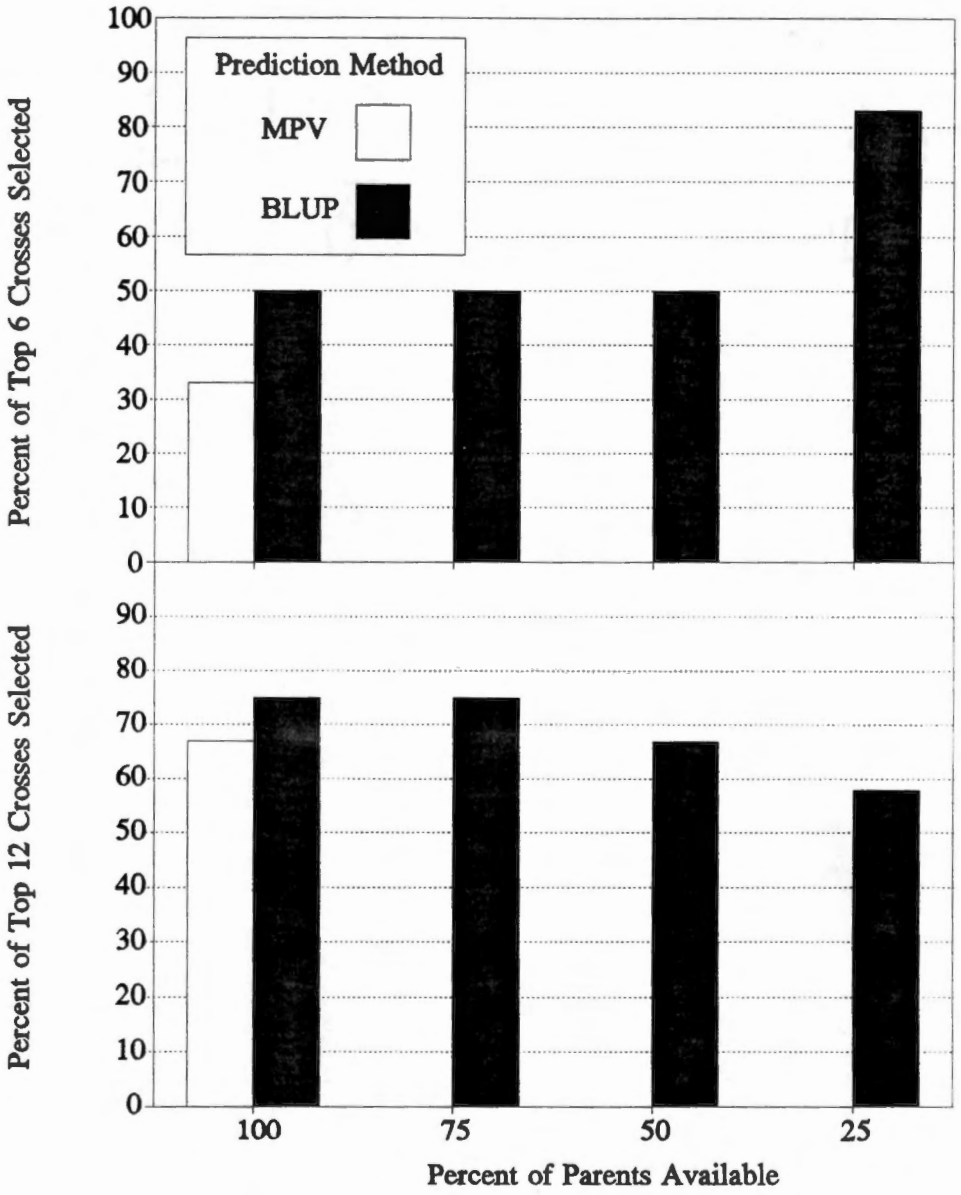


Fig. 1. Percentages of the top 12 and top six crosses overall selected from predictions made from 100, 75, 50, and 25% of the 25 parents in this study using MPV (at 100% only) and BLUP.

PART III

Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

II. Influence of Performance of Relatives on Breeding Values of Parents

Abstract

The best estimate of the performance of a future progeny is the mean of the performance of its parents. In order for progeny predictions to be accurate, the estimates of the performance of each parent, or breeding value, must first be precise. In cases where data for a parent is unavailable, the alternative is to estimate its breeding value based on evaluations of its close relatives, preferably its progeny if such data are available. Best linear unbiased predictions (BLUP) have been shown to provide fairly precise predictions of the breeding values of dairy cattle sires for milk yield based on information from their dams and daughters. The effects of progeny information on the breeding values of their parents using BLUP has not been established in plant breeding. The objectives of this study were to i) compare the effects of soybean (*Glycine max* (L.) Merr.) progeny and grand-progeny yield performance on the BLUP of predicted breeding values of their parents, and ii) determine how much information on a parent *per se* is required before progeny information no longer affects the predicted breeding value of the parent. The F_4-F_6 bulks and $F_{5:6}$ lines from 24 soybean crosses, two major parents, and one major grandparent were evaluated in 11 different environments in Tennessee. Six and seven of the 24 crosses were progeny from two of the major parents, Hutcheson and TN 4-86, respectively; and 11 crosses were grand-progeny of Essex. The BLUP of breeding values for each of the three parents were calculated using varying numbers

of random records on the parent *per se* and on the progeny or grand-progeny of the parent. Predicted breeding values and performance of future crosses based on the breeding values of the parents were compared to the actual values obtained from the field evaluations. Adding performance records on progeny had little effect on the predicted breeding values of Hutcheson and TN 4-86 when there were 10 observations on the parent *per se*; and the breeding value of Essex was affected much less by records from its grand-progeny than by records from Essex *per se*. These results indicate that information from relatives has insignificant effects when adequate data are available from the parents *per se*. Five records on Hutcheson *per se* plus five random progeny from Hutcheson was equal to 10 records on Hutcheson *per se* in predicting all of the top six crosses; and five records on TN 4-86 *per se* plus five random progeny from it progeny was equal to 10 records on TN 4-86 *per se* in predicting the bottom four crosses. These results illustrate how progeny performance can be utilized by BLUP as "substitute" information in predicting breeding values of parents.

CHAPTER 1

Introduction and Literature Review

For traits which are under primarily additive genetic control, the best estimate of the performance of a future progeny is the mean of the performance of its potential parents. Therefore, to predict progeny performance accurately, the best possible estimates of the performance, or breeding values, of its parents are required. A typical method of estimating performance of parents is to use a fixed linear model and calculated best linear unbiased estimates (BLUE) for each parent of interest. One problem many plant breeders face is that there is either limited or no performance data on the parents of interest, thus making accurate predictions of the performance of future progeny very difficult to obtain. In such cases, the alternative is to evaluate progeny of a parent, if such data are available, and estimate the breeding value of a parent based on the performance of its progeny.

Henderson (1975a; 1977) described the use of a mixed linear model to calculate the best linear unbiased predictions (BLUP) of breeding values of potential parents based on observed records and known (or estimated) variance/covariance structure among random effects. Henderson's methods utilize the genetic

relationship among individuals as the variance/covariance among random genetic effects, and assume that the only cause of correlation between observations on different individuals is additive genetic variance (1975b). By utilizing the genetic variance/covariance among individuals, individuals contribute to the predicted values of their relatives. He demonstrated later how a narrow sense estimate of heritability of the trait of interest also could be incorporated into the model (1977). In cases of balanced data and no correlation among genotypes, the mixed model solutions (BLUP) and least squares solutions, best linear unbiased estimates (BLUE), lead to the same relative ranking of genotypes (White *et al.*, 1986).

One important property of BLUP is that performance predictions of individuals are supplemented by performance of its relatives and the effect is determined by the magnitude of the relationships between the individual and its relatives. This feature of BLUP has been demonstrated in dairy cattle (Henderson, 1977) and soybean (Allen *et al.*, 1991). Another prominent feature of BLUP is that it "shrinks" the prediction of an individual as a function of the effective number of observations on the individual. Hill and Rosenberger (1985) illustrated the effectiveness of the shrinkage estimator by combining unbalanced data from mean yields of alfalfa (*Medicago sativa* L.) genotypes from a series of evaluations trials conducted at one location over a period of eight years. Allen *et al.* (1991) and Panter and Allen (1991) demonstrated the activity of the shrinkage estimator in predicting yield of future soybean crosses.

Several studies on the use of BLUP in plant breeding applications have been reported. Allen and Bhardwaj (1987) utilized BLUP to evaluate the predicted *versus* actual yield response of soybean ancestral lines and cultivars to simulated acid rain. Allen *et al.* (1991) used simulated data to demonstrate, conceptually, how BLUP could be used to predict parental breeding values of soybean strains in cases of balanced and unbalanced parental data, and how BLUP compared to traditional midparent values (MPV). In Part II, I showed that the mean performance and relative rankings of 24 soybean crosses could be predicted from historical records on their parents more precisely using BLUP than from traditional MPV.

Although the properties of BLUP are well documented, the magnitude of the effect of progeny information on breeding values of their parents has not been demonstrated sufficiently in plant breeding applications. Another question which needs to be addressed is how much information on a parent *per se* is required before information on its relatives no longer contributes significantly to its breeding value. Suitable answers to these questions could provide critical information to plant breeders who wish to utilize BLUP as a method for estimating breeding values of potential parents.

The objectives of this study were to determine i) the effects of progeny and grand-progeny on the BLUP of breeding values of parents, and ii) how much parental information is necessary before information on relatives no longer affects parental breeding values.

CHAPTER 2

Materials and Methods

Twenty-four unselected F_4 soybean crosses, representing diverse genetic backgrounds, plus two major parents ('Hutcheson' and 'TN 4-86') and one major grand-parent ('Essex') of the 24 crosses, were chosen as entries for this study (Table 1[2]). The 24 crosses were not predetermined for this study, but were a sample of a larger group of crosses processed yearly in the Univ. of Tennessee Soybean Breeding Program (UTSB). The single criterion for selection was that there be enough seed available for a cross to be entered into a replicated yield trial. The crosses were derived as described in Part II.

Field Performance Trials

The 27 entries were evaluated for seed yield as described in Part II. Parent plots were four rows of the same parent. In 1990, the two center rows of parent

[2] Tables and figures are presented in the Appendix.

plots were harvested and considered as one observation, and one border was harvested and considered as a separate observation in the same environment in 1990. This scheme provided equal numbers of observations for the crosses (1 bulk and 1 line mean) and parents (2) in each 1990 environment. All yields were adjusted to 130 g H_20 kg^{-1}. These yield trials had two purposes. The first purpose was to provide multiple observations on both lines and bulks of the 24 crosses across a broad range of environments which could be used to predict the breeding values of their parents. The second purpose was to provide multiple observations of the three parents. These records would be used as target breeding values to evaluate predictions made from observations on the progeny, and to simulate historical data which may be available to a breeder.

Statistical Analyses

Each location/year combination was considered to be a single environment. Overall least squares mean yields were calculated for both bulk (11 environments) and lines (5 environments) for all 24 crosses. Blocks were considered random and all other effects were considered fixed. Using rank correlation analysis, it was determined that bulk and line data could be combined (for a total of 16 records on each cross) to predict the breeding values of the parents. Overall least squares means for the three parents were calculated across environments (total of 16 records)

and these estimates were considered to be the target breeding values for later analyses.

Effects of Relatives on Breeding Value of Parents. For each of two major parents, TN 4-86 and Hutcheson, a series of predictions was made using BLUP from a data set consisting of the mean yield for each of the 24 crosses at each of the 11 environments. For each parent, predictions were made by assuming no data available (0 records) *versus* 5 or 10 random historical records for the parent, and supplementing the parental data with no progeny data (0) or 5, 10, 20, or 50 on random progeny of the parent. Random historical records on the parents were simulated by using the parental data collected in the field trials (1988-1990). These combinations of parent and progeny data were considered to be typical of the kinds of data a breeder might possess for estimating breeding values of potential parents. The purpose of these analyses was to determine the effect of progeny information on the breeding values of parents using BLUP.

The mixed linear model was:

$$y = \mu + X\beta + ZU + \epsilon ,$$

where

y is a vector of measured yield,

μ is the overall mean yield,

X is a design matrix of fixed environments,

ß is the vector of unknown effects due to environments to be estimated,

Z is a design matrix of random genotypes,

U is the vector of unknown effects due to genotypes, and

ϵ is an error vector with a null mean.

Since BLUP required an estimate of heritability for the given trait, variance components were estimated from 1990 $F_{5:6}$ line data using the REML option in SAS® PROC VARCOMP (SAS, 1985) and heritability for yield (0.24, data not shown) was calculated based on these estimates. The heritability estimate plus the genetic relationship matrix between each parent and the 24 crosses, which represents the variance/covariance between random effects, were incorporated into the mixed model equations as described by Allen *et al.* (1991).

To determine the precision of predictions, the standard errors of the differences between the actual and the predicted values ($s_{\bar{d}}$) for each parent were calculated for each series of predictions. When predicting breeding values of parents, the actual predictions (or error of the predictions) is of limited importance, but the variances of the errors of the predictions are usually critical. That is, a plant breeder chooses parents based on the ranks of their breeding values and not on the predicted values themselves. The $s_{\bar{d}}$ was therefore considered the superior statistic for comparing sets of predicted breeding values. Comparing the $s_{\bar{d}}$ between sets of predictions also provided a means to determine the effect of progeny information on the breeding values of the parents. Changes in the precision of the predictions could

be attributed directly to additional information accumulated when progeny records were added to the predictive data set.

Overall mean seed yield from the actual data revealed that the top six crosses (CROSS-22, CROSS-23, CROSS-10, CROSS-24, CROSS-04, and CROSS-18) were ones which had Hutcheson as one of the parents, and four of the bottom six crosses (CROSS-13, CROSS-08, CROSS-01, and CROSS-17) were ones which had TN 4-86 as one of the parents (Table 2). Therefore, counts were made to ascertain how many of the top six crosses would have been predicted as such based on the BLUP breeding value of Hutcheson, and how many of the four low-yielding TN 4-86 crosses were among the predicted bottom six crosses based on the BLUP breeding value of TN 4-86.

In order to determine the effects of more-distantly related progeny on the breeding values of potential parents, the same suite of predictive analyses was conducted as described above except that Essex (a parent of Hutcheson and grandparent of 11 of the 24 crosses, Table 2), was the parent of interest. The pedigree diagram in Fig. 1 shows an example of one of the 11 crosses, CROSS-04 (TN 4-86 x Hutcheson), which was a grand-progeny of Essex. Predictions were made by assuming no data available (0 records) *versus* 5 or 10 random historical records for Essex, and supplementing the parental data with no grand-progeny data (0) or 5, 10, 20, or 50 on random grand-progeny of Essex. Historical data were simulated by using records collected in the field trials (1988-1990). The $s_{\bar{d}}$ between the actual and the predicted values for Essex were calculated to determine the

precision of each set of predictions and how much effect the information on grand-progeny had on the breeding value of Essex.

CHAPTER 3

Results and Discussion

Actual Yield Performance of Parents and Progeny

Rank correlation between least square mean seed yields across environments for F_4-F_6 bulk and $F_{5:6}$ line data was found to be positive (r_s=0.77) and highly significant ($P \leq 0.01$). Based on this observation, it was determined that bulk and line data were equally accurate indicators of the overall cross mean and therefore could be combined to predict the breeding values of Hutcheson, TN 4-86, and Essex. Combining these data provided 16 observations (11 bulk and 5 line means) across 11 environments for each cross.

The best measures of the breeding values (least squares means) of Hutcheson, TN 4-86, and Essex from the same 11 environments were 2767, 2012, and 2262 kg ha^{-1}, respectively (Table 2). Since the top six crosses (overall ranks 1-6, Table 2) were direct progeny of Hutcheson and four of the bottom six crosses (ranks 19-24) were direct progeny of TN 4-86, a superior method of predicting the breeding values

of Hutcheson and TN 4-86 should also predict, with greater precision, these top six and bottom four crosses.

Effects of Progeny Information on Breeding Values of Parents

Example of a High-Yielding Parent. When there were no records on Hutcheson *per se*, the predicted value for Hutcheson increased as the number of records on its progeny increased (2143 to 2496 kg ha^{-1}, Table 3). The steady increase in these predictions was due to the shrinkage estimator and was in agreement with observations from other workers (Hill and Rosenberger, 1985; Allen *et al.*, 1991). When observations on a small number of progeny (*e.g.* 5) of Hutcheson were available, the predicted value for Hutcheson was adjusted back toward the overall mean (16 kg ha^{-1} > overall mean, Table 3); as observations on more progeny (*e.g.* 50) became available, the predicted value for Hutcheson was adjusted less toward the overall mean and more toward the mean of its progeny (278 kg ha^{-1} > overall mean, Table 3). Since none of the progeny of Hutcheson yielded as much as Hutcheson, the predicted breeding value of Hutcheson approached, but never equalled, its actual yield (2767 kg ha^{-1} or 502 kg ha^{-1} > overall mean, Table 3).

Another important point is that even when there were no data for either Hutcheson or its progeny, BLUP could still predict a breeding value for Hutcheson

(2143 kg ha^{-1}, Table 3). This value was calculated from the progeny performance of parents to which Hutcheson was related. For example, of the 18 crosses which were not direct progeny of Hutcheson, seven had covariance coefficients with Hutcheson of 0.18 or greater, and all 18 had at least some genetic covariance with Hutcheson (Table 1). In the absence of actual observations on Hutcheson or its direct progeny, all of the other 18 crosses contributed to the predicted breeding value of Hutcheson, with the degree of contribution determined by the magnitude of their genetic relationships with Hutcheson.

These results demonstrate two important properties of BLUP in genetic applications. First, observations on relatives of an individual contribute to the predicted value of the individual. Second, as more observations on relatives become available, the *effective* number of observations on the individual increases. The effective number of observations is determined by the number of relatives and the degree of their relationships with the individual; therefore, one observation on a relative is not equivalent to one observation on the individual *per se*.

As expected, $s_{\bar{d}}$ between the actual and the predicted breeding values of Hutcheson decreased as the number of actual observations on Hutcheson went from zero to 10 (73 to 19 kg ha^{-1}, Table 3). These results indicate that the variance of prediction errors decreased as more records on Hutcheson (*i.e. per se* and via progeny) became available.

The ability of BLUP to identify the top six crosses using varying numbers of records on Hutcheson *per se* and 0 and 5 records on its progeny is presented in

Fig. 2. When there were no records on progeny of Hutcheson, BLUP identified up to 100% of the top six crosses solely from records on Hutcheson. Five records on Hutcheson *per se* plus five random progeny from Hutcheson was equal to 10 records on Hutcheson *per se* in identifying all of the top six crosses. These results illustrate how progeny performance can be utilized by BLUP as "substitute" information in predicting breeding values of parents.

Example of a Low Yielding Parent. The actual and the predicted breeding values of TN 4-86 with different numbers of records on TN 4-86 *per se* and its progeny are presented in Table 4. The predicted breeding value of TN 4-86 tended to fluctuate both positively and negatively as the number of records on progeny of TN 4-86 increased. This movement in both directions probably was caused by the random inclusion of CROSS-04, a high-yielding progeny of TN 4-86 (Table 2), in the predictive data set. When there were no data for either TN 4-86 or its progeny, the predicted breeding value of TN 4-86 was 2204 kg ha^{-1} (Table 4). This prediction was derived from observations on the other 18 crosses which were related to TN 4-86. Examination of the genetic covariances showed that eight crosses had covariance coefficients with TN 4-86 of 0.25 or greater (Table 2). In particular, three crosses, CROSS-09, CROSS-15, and CROSS-20, had covariances of at least 0.31 with TN 4-86. A relationship of this magnitude is approximately equivalent to the relationship between an individual and its grandparents. The effect of information from these relatives was illustrated by the low $s_{\bar{d}}$ in all cases (Table 4).

The degree of the relationships between TN 4-86 and the 24 crosses in this study contributed to the precision of the predicted breeding values of TN 4-86.

The BLUP method of prediction identified up to 75% of the four low-yielding crosses with TN 4-86 when there were no data on the progeny of TN 4-86 (Fig. 3). However, as more progeny were included in the predictive data sets, fewer of the four crosses were identified as low-yielding progeny. These observations also demonstrate the effect of progeny performance information of the breeding values of parents. Since TN 4-86 was a parent of three crosses (CROSS-04, CROSS-03, and CROSS-11) which had medium to high yield performance, the predicted breeding value of TN 4-86 increased. The effect of this increase was a corresponding increase in the predicted performances of the four low-yielding crosses.

Effect of Grand-Progeny on the Breeding Value of Essex

When there were no data for Essex or any of its grand-progeny, BLUP predicted the breeding value of Essex to be 2193 kg ha^{-1}, which was only 69 kg ha^{-1} less than the actual value (2262 kg ha^{-1}, Table 5). This prediction was based on performance records of the 13 crosses which were not grand-progeny of Essex (Table 1). Notably, two of those crosses, CROSS-06 and CROSS-16, had genetic covariances with Essex of 0.34, which is roughly equivalent to the genetic

covariances of Essex and it grand-progeny. Since each of those crosses had 16 records, the breeding value of Essex was predicted with the *equivalent* of 32 records on its grand-progeny. Therefore, as records on the true grand-progeny were added to the predictive data set, their effect was minimal the variance and corresponding $s_{\bar{d}}$ among predictions was small (21 kg ha^{-1}, Table 5).

When five or 10 records on Essex were available, the $s_{\bar{d}}$ was greater than when there were no records on Essex (26 and 23 *versus* 21 kg ha^{-1}, respectively). The increase in the variance of predictions was due to the strong effect of records on Essex *per se* compared to records from its grand-progeny and the two crosses highly related to Essex. These cases illustrated that information on an individual *per se* had a much greater impact on its predicted breeding value than information on its distant relatives.

The results from these analyses indicated that grand-progeny indeed affect the BLUP of the breeding values of their grandparents, but to a much lesser degree than direct progeny on their parents. When there were no data for an individual, using sufficient numbers of records of its grand-progeny provided a fairly precise prediction of its breeding value. When a few records were available for the individual, the effects of information from its grand-progeny were limited relative to the information on the individual *per se*.

CHAPTER 4

Conclusions

The advantage of using BLUP to predict breeding values of potential parents is greatest when data on the parents are either unavailable or very limited, but data on progeny and/or close relatives of potential parents are available. Under these conditions, the shrinkage property of BLUP helps provide a conservative estimate of potential parents. When more data are accumulated on these parents, the effects due to information on their relatives decreases and the predicted breeding values are adjusted closer to the observed mean. We concur with the observations of Hill and Rosenberger (1985) that the shrinkage estimator is a very desirable property in terms of predicting the least biased value of a genotype for a given situation.

In the absence of information for a parent *per se*, performance information from its close relatives provide only a moderately precise prediction of its breeding value and the performance level of potential crosses with the it. When performance information on the parent *per se* is added to information from its relatives, prediction of the performance level of future crosses with the parent is greatly enhanced; that is cross combinations with the parent which result in superior and/or inferior progeny

can be predicted much more accurately when information from other relatives (which share some of the same genes) is available. These properties of BLUP should prove to be very valuable to plant breeders. In situations where data are limited on a potential parent, performance data from its close relatives can provide the necessary information to make better decisions on paired matings with the parent.

Results from these analyses of the soybean crosses in this study also suggest that there is a point when accumulating more records on relatives of an individual has little or no effect on the predicted breeding value of the individual. When 10 to 20 observations have been recorded for an individual, adding records on relatives more distantly related than progeny or half-sibs has negligible effects on the predictions of the individual. When more than 20 observations for an individual are available, adding information from relatives as closely related as full-sibs has little effect on the predicted breeding value of the individual. These observations should provide guidance to breeders who wish to use BLUP but are unsure of how much information should be maintained in order to make accurate predictions of breeding values of potential parents.

Literature Cited

Allen, F.L. and H.L. Bhardwaj. 1987. Genetic relationships and selected pedigree diagrams of North American soybean cultivars. Tennessee Agric. Exp. Sta. Bull. 652.

Allen, F.L., D.M. Panter, W.L. Sanders, and R.A. McLean. 1991. Best linear unbiased prediction as a method to enhance breeding for yield in soybeans. Crop Sci. (in review).

Henderson, C.R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-477.

Henderson, C.R. 1975b. Use of all relatives in intraherd prediction of breeding values and producing abilities. J. Dairy Sci. 58:1910-1916.

Henderson, C.R. 1977. Prediction of future records. p. 615-638. *In* E. Pollack, *et al.* (ed.). Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames, IA.

Hill, R.R. and Rosenberger, J.L. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470.

SAS Institute Inc. 1985. SAS® Users's Guide: Statistics, Version 5 Edition. Cary, NC. SAS Institute Inc. 956 pp.

White, T.L., G.R. Hodge, and M.A. Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. p. 99-122. *In* Workshop of the Genetics and Breeding of Southern Forest Trees RIEG, June 25-26, 1986, Univ. of Florida, Gainsville, FL.

Appendix

Table 1. Pedigrees, and genetic variance†/covariance‡ among two major parents, TN 4-86, and Hutcheson, and a grandparent, Essex, and the 24 crosses in this study.

Parent or Cross I.D.	Pedigree	Parent		
		Essex	TN 4-86	Hutcheson
Parents				
Essex	Lee/S5-7075	1.03†		
TN4-86	Bedford/Crawford	0.10‡	1.01†	
Hutcheson	V68-1034/Essex	0.54	0.06‡	1.02†
Crosses				
CROSS-01	D83-3349/TN4-86	0.14	0.75	0.08‡
CROSS-02§	D83-3349/V81-141	0.37	0.28	0.20
CROSS-03	Stafford/TN4-86	0.06	0.56	0.07
CROSS-04§	TN4-86/Hutcheson	0.32	0.54	0.54
CROSS-05	TN82-221/TN84-7	0.24	0.25	0.13
CROSS-06	TN84-3/TN85-157	0.34	0.19	0.19
CROSS-07	TN84-7/TN85-55	0.21	0.25	0.11
CROSS-08	TN85-10/TN4-86	0.10	0.66	0.06
CROSS-09	TN85-102/TN84-21	0.18	0.31	0.10
CROSS-10§	TN85-116/Hutcheson	0.54	0.08	0.66
CROSS-11§	TN85-117/TN4-86	0.33	0.55	0.18
CROSS-12§	TN85-121/TN85-116	0.36	0.13	0.20
CROSS-13§	TN85-13/TN4-86	0.33	0.55	0.18

Table 1 (continued)

Parent or Cross I.D.	Pedigree	Parent		
		Essex	TN 4-86	Hutcheson
CROSS-14	TN85-136/TN85-55	0.05	0.16	0.03
CROSS-15	TN85-162/TN85-172	0.20	0.31	0.11
CROSS-16	TN85-172/TN85-157	0.34	0.16	0.18
CROSS-17	TN85-22/TN4-86	0.07	0.53	0.04
CROSS-18§	TN85-3/Hutcheson	0.31	0.10	0.53
CROSS-19§	TN85-55/TN5-85	0.34	0.23	0.18
CROSS-20	TN85-55/TN83-26	0.08	0.33	0.05
CROSS-21	TN85-55/TN84-3	0.21	0.25	0.11
CROSS-22§	Hutcheson/TN82-221	0.33	0.18	0.54
CROSS-23§	Hutcheson/TN85-55	0.30	0.17	0.53
CROSS-24§	D83-3318/Hutcheson	0.36	0.27	0.56

§ denotes crosses which were grand-progeny of Essex

Table 2. Overall mean yields and ranks of the three parents and 24 soybean crosses (means of F_4, F_5, and F_6 bulks, and means of $F_{5:6}$ lines) evaluated in yield trials, 1988-1990.

Parent	Yield			Rank		
or Cross	Bulk†	Lines‡	Overall§	Bulk	Lines	Overall
	------ kg ha^{-1} ------					
Parents						
Hutcheson	2843	2028	2767			
Essex	2295	1604	2262			
TN4-86	2021	1408	2012			
Crosses						
CROSS-22¶	2645	1882	2589	2	2	1
CROSS-23¶	2657	1838	2584	1	3	2
CROSS-10¶	2489	1983	2514	4	1	3
CROSS-24¶	2548	1828	2506	3	4	4
CROSS-04¶	2372	1806	2378	5	5	5
CROSS-18¶	2347	1643	2310	6	8	6
CROSS-06	2310	1651	2287	7	7	7
CROSS-09	2286	1663	2274	8	6	8
CROSS-21	2282	1633	2262	9	9	9
CROSS-16	2254	1574	2224	10	15	10
CROSS-15	2194	1628	2200	13	10	11
CROSS-19	2237	1502	2190	11	18	12
CROSS-03	2127	1604	2146	17	11	13
CROSS-14	2150	1537	2141	15	17	14
CROSS-12	2115	1596	2135	18	12	15
CROSS-07	2212	1377	2133	12	22	16
CROSS-11	2104	1579	2123	19	14	17
CROSS-20	2164	1439	2120	14	19	18
CROSS-13#	2077	1564	2099	20	16	19
CROSS-05	2145	1350	2079	16	23	20

Table 2 (continued)

Parent or Cross	Yield			Rank		
	Bulk†	Lines‡	Overall§	Bulk	Lines	Overall
	------ kg ha^{-1} ------					
CROSS-08#	2071	1432	2054	21	20	21
CROSS-01#	2005	1401	1999	22	21	22
CROSS-17#	1846	1585	1947	24	13	23
CROSS-02	1966	1311	1944	23	24	24

† Least squares means of three replications at one location in 1988 and five locations in 1989 and 1990 (11 environments).
‡ Least squares means of 40 lines (eight lines at each of five locations (environments)) in 1990.
§ Least squares means of bulks at 11 environments (1988-1990) and lines at five environments (1990).
¶ indicates high-yielding progeny of Hutcheson.
# indicates low-yielding progeny of TN 4-86.

Table 3. Actual and predicted breeding values, and standard errors of the differences ($s_{\bar{d}}$ = actual *versus* predicted) of breeding values of Hutcheson with varying numbers of records on Hutcheson *per se* and progeny of Hutcheson.

No. Records		Predicted					
		No. of Records on Progeny of Hutcheson					
Hutcheson	Actual†	0	5	10	20	50	$s_{\bar{d}}$
		---------- kg ha^{-1} ----------					
0	2767 (502)‡	2143 (16)§	2293 (127)	2311 (151)	2542 (339)	2496 (278)	73
5		2412 (237)	2601 (385)	2744 (501)	2616 (388)	2672 (440)	55
10		2715 (488)	2688 (466)	2604 (395)	2641 (424)	2661 (437)	19

† Least squares means across 11 environments (1988-1990).
‡ Difference between actual value and the overall mean.
§ Difference between predicted value and the overall mean.

Table 4. Actual and predicted breeding values, and standard errors of the differences ($s_{\bar{d}}$ = actual *versus* predicted) of breeding values of TN 4-86 with varying numbers of records on TN 4-86 *per se* and progeny of TN 4-86.

		Predicted					
No. Records		No. of Records on Progeny of TN 4-86					
TN 4-86	Actual†	0	5	10	20	50	$s_{\bar{d}}$
		---------- kg ha^{-1} ----------					
0	2012	2204	2254	2094	2117	2085	33
5		1974	2105	1965	1992	2068	27
10		2144	2061	2073	2070	1999	23

† Least squares means across 11 environments (1988-1990).

Table 5. Actual and predicted breeding values, and standard errors of the differences ($s_{\bar{d}}$ = actual *versus* predicted) of breeding values of Essex with varying numbers of records on Essex *per se* and grand-progeny of Essex.

No. Records		Predicted					
		No. Records on Grand-progeny of Essex					
Essex	Actual†	0	5	10	20	50	$s_{\bar{d}}$
		---------- kg ha^{-1} ----------					
0	2262	2193	2259	2270	2234	2318	21
5		2188	2196	2322	2207	2280	26
10		2314	2202	2325	2294	2250	23

† Least squares means across 11 environments (1988-1990).

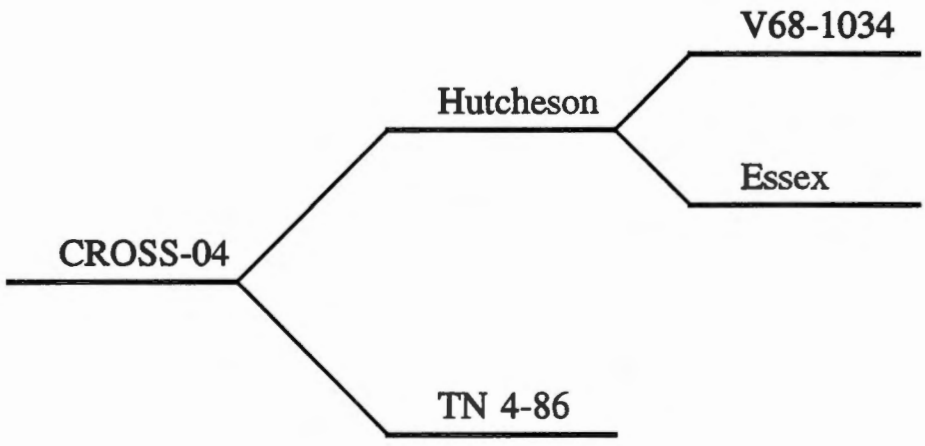


Fig. 1. Pedigree diagram showing the relationships between CROSS-04, Hutcheson, and Essex.

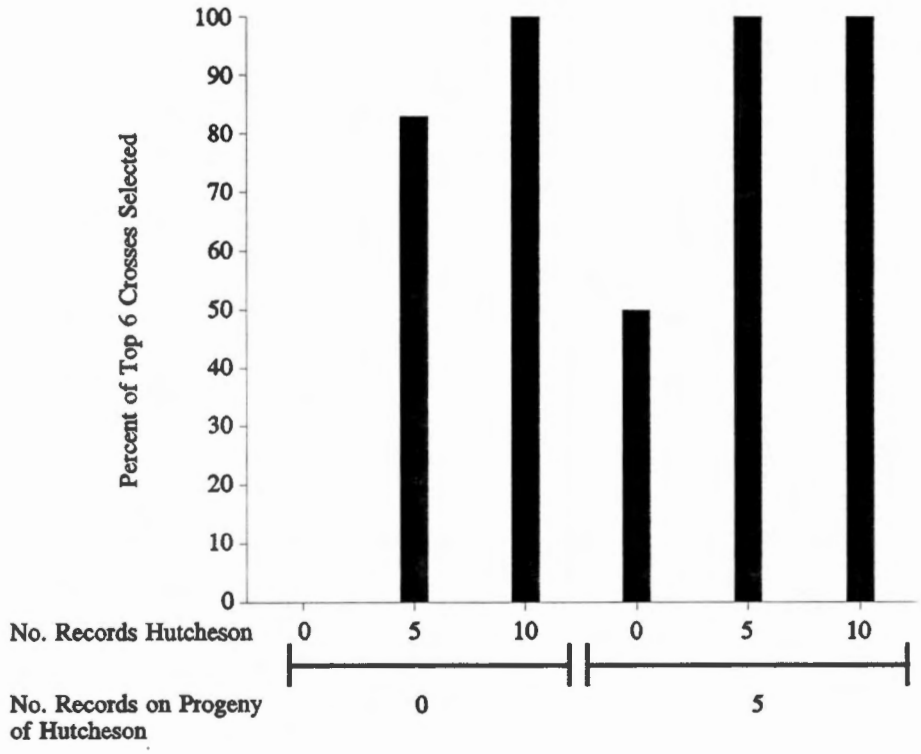


Fig. 2. Percentages of the top six crosses overall selected based on the predicted breeding value of Hutcheson with varying numbers of records on Hutcheson and its progeny.

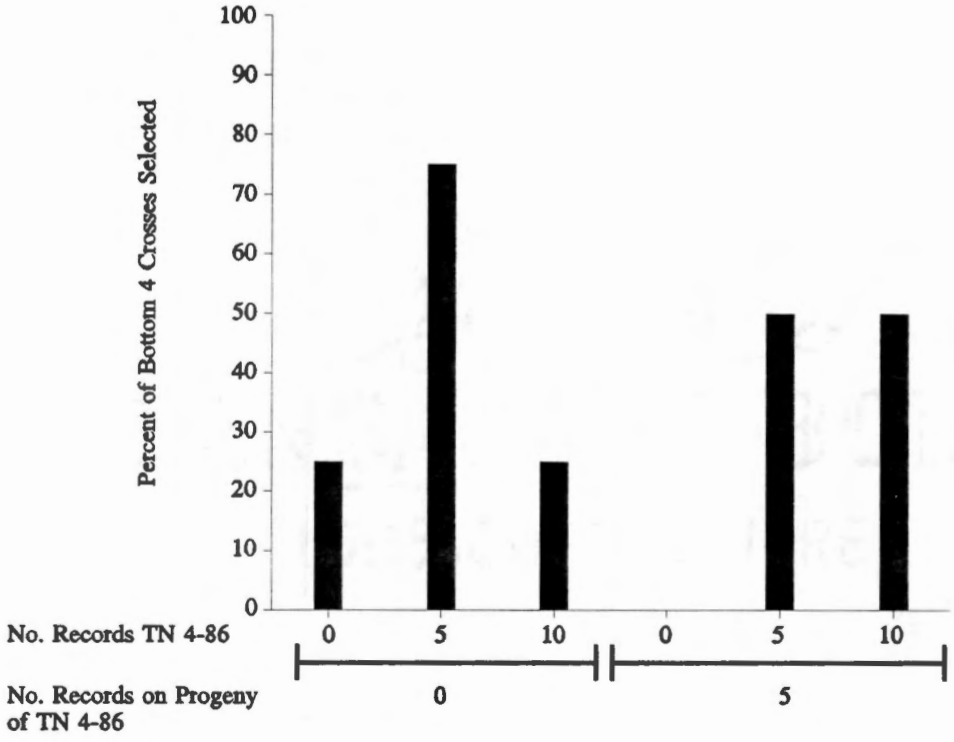


Fig. 3. Percentages of the bottom four crosses (derived from TN 4-86) overall selected based on the predicted breeding value of TN 4-86 with varying numbers of records on TN 4-86 and its progeny.

PART IV

Using Mixed Linear Models and Best Linear Unbiased Predictions to Predict Seed Yield in Soybeans:

III. Selection of Superior Progeny from a Limited Number of Yield Trials

Abstract

Performance testing new genetic material is one of the most important and expensive facets of a plant breeding program. An accurate estimate of the performance of a new line is critical for selection and is usually obtained from a series of performance trials across several years and/or locations. Economic constraints on many plant breeding programs have forced breeders to limit the numbers of years and/or locations for performance testing, and new statistical approaches are needed which maximize the accuracy of estimates of performance from few environments. Best linear unbiased predictions (BLUP) have been shown to be superior predictors of breeding values of potential parents in both dairy cattle and soybeans. This methodology, which augments predictions of individuals with little or no performance information by using observations on their close relatives, should be applicable to performance testing restricted to few environments. The objectives of this study were to determine i) if BLUP were more accurate estimators of genotypic mean performance than the common approach of best linear unbiased estimations (BLUE) from a series of soybean (*Glycine max* (L.) Merr.) performance evaluations conducted in one or two environments, and ii) if the inclusion of historical parental data sufficiently enhanced the accuracy of BLUP to justify maintaining parental information for future use. Bulks and lines of 24 soybean crosses and four check cultivars were evaluated at 11 different environments in Tennessee to estimate the

mean seed yield performance of each cross and cultivar. Historical records on the yield performance of the parents of each cross were compiled from yield trials conducted in Tennessee from 1982-1990. Data from three Tennessee locations were used in various combinations of one location in one year, one location across two years, and two locations in one year to predict mean seed yield performance of the 28 entries using three methods: 1) BLUE, 2) BLUP without parental data (BLUP(NP)), and 3) BLUP with parental data included (BLUP(P)). Standard errors of the differences ($s_{\bar{d}}$) and rank correlations (r_s) between the actual mean yields calculated across all environments and the predicted mean yields showed that either method of BLUP was superior to BLUE for improving the precision of the yield estimates. Generally, there was no difference between BLUP(NP) and BLUP(P), indicating that adding historical parental information did not increase the accuracy of BLUP for the genetic material used in this study. Had the material used in this study been more genetically diverse, BLUP(P) would have probably been superior to BLUP(NP) for improving the precision of the yield estimates.

CHAPTER 1

Introduction and Literature Review

Performance testing new genetic material is one of the most important and expensive facets of most plant breeding programs. Selecting superior lines from among a large group of lines is usually accomplished by evaluating the group across several locations and years. The number of different environments used for these evaluations is dictated generally by the economic resources available. Unfortunately, economic constraints on many plant breeding programs across the country have forced breeders to limit the number of environments used for evaluating and selecting superior genetic material. It has therefore become imperative that breeders investigate new statistical approaches that may maximize the accuracy of the estimate of the performance of a line from fewer environments.

The classical approach to estimating genotypic means has been to calculate a simple arithmetic mean across environments and to make selections based on these means. However, in cases where all genotypes were not evaluated or data were missing in some environments, performance estimates were biased by the number and kinds of environments in which they occurred. An improvement over arithmetic

means was to consider all effects fixed and to calculate the least squares or best linear unbiased estimate (BLUE) for each genotype. When data were not balanced, means were adjusted to reflect the estimated performance as though each genotype had been evaluated in each environment. If the data were balanced, BLUE of each genotype was equal to its arithmetic mean. The BLUE values for a set of genotypes across a range of environments can be easily calculated using SAS® PROC GLM (SAS, 1985).

A further refinement to estimating genotypic means was to use a mixed linear model and calculate the best linear unbiased prediction (BLUP) for each genotype. Henderson (1952, 1963) introduced most of the concepts of BLUP in the form of a selection index. He demonstrated how BLUP could be calculated by any solution of the equations:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} * \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

where X and Z are design matrices of fixed and random effects, respectively; R and G are estimates of the variances of error and random effects, respectively; and b and u are the unknown vectors of fixed and random effects to be estimated. Until recently, Henderson's methodology has found only limited acceptance in the plant breeding community. Hill and Rosenberger (1985) demonstrated the superiority of BLUP over other methods of combining unbalanced data for estimating mean yields of alfalfa (*Medicago sativa* L.) genotypes from a series of evaluations conducted at one location over a period of eight years. Bridges (1989) showed similar results from mixed models analyses in estimating mean yields of four cucumber (*Cucumis*

sativus L.) cultivars at two locations using a latin square design. In both of these studies, the genotypes evaluated were assumed to be unrelated.

Perhaps a primary reason for the lack of use of Henderson's mixed model methodology has been the difficulty in computing solutions to the mixed model equations. However, computer programs written in SAS/IML™ (SAS, 1988) which perform various types of mixed model analyses have been published (Blouin *et al.*, 1989; Bridges, 1989; McLean, 1989; Sanders, 1989; and Stroup, 1989). Blouin *et al.* (1989) developed a computer program, GLMM, which virtually eliminates the need to write computer code to perform most types of mixed model analyses. Currently, GLMM is restricted to analyses where no covariances exist between random effects and therefore genotypes must be considered to be unrelated.

Henderson (1975a; 1977) described the genetic application of mixed linear models to calculate the BLUP of breeding values of potential dairy cattle sires based on observed records of their dams and daughters, and known (or estimated) variance/covariance structure among sires and potential female mates. Henderson's methods utilize the genetic relationship among individuals as the variance/covariance among genetic effects and assume that the only cause of correlation between observations on different individuals is additive genetic variance (1975b). By utilizing the genetic variance/covariance among individuals as the G matrix, related individuals contribute to the predicted values of one another. He demonstrated later (1977) how a narrow sense estimate of heritability of the trait of interest could be incorporated into the model. In cases of balanced data and no correlation among genotypes, the

mixed model solutions (BLUP) and least squares solutions (BLUE) lead to the same relative ranking of genotypes (White *et al.*, 1986).

Allen and Bhardwaj (1987) incorporated genetic relationships in mixed models equations to evaluate the actual *versus* predicted yield response of soybean ancestral lines and cultivars to simulated acid rain. Allen *et al.* (1991) used simulated data to illustrate the potential effectiveness of BLUP in predicting yield performance of future soybean genotypes. In Part II of this dissertation, I demonstrated the superiority of BLUP to classical midparent values for predicting yield of future soybean crosses from historical records on their parents, and in Part III, I described the effects of records on relatives on the breeding values of parents using BLUP.

It appears that the incorporation of the genetic relationships among soybean lines evaluated for yield in a limited number of environments may improve the precision of the estimated yield performance of the individuals. If these estimates are more precise than classical approaches to genotypic mean estimation, breeders should have more confidence in the selection decisions made based on these estimates. Since one of the properties of BLUP is that related individuals contribute to the predictions of one another, it seems logical that including records on parents or close relatives of the genotypes being evaluated should improve both the precision and the accuracy of the adjusted yield values by increasing the effective number of observations on each genotype.

The objectives of this study were to determine i) if BLUP values were more precise estimators of genotypic mean yield performance than BLUE from trials

conducted in one or two environments, and ii) if the inclusion of historical parental data sufficiently enhanced the precision of BLUP to justify maintaining parental yield information for the sole purpose of augmenting predictions of their progeny using BLUP.

CHAPTER 2

Materials and Methods

Twenty-four unselected F_4 soybean crosses, representing diverse genetic backgrounds, plus four adapted check cultivars, Hutcheson, TN 4-86, TN 5-85, and Essex, were chosen as entries for this study (Table 1[3]). The 24 crosses were not predetermined for the purpose of this study, but were chosen from a larger set of crosses that had been made as a regular part of the cultivar development efforts of the Univ. of Tennessee Soybean Breeding Program (UTSB). The single selection criterion was that there be enough seed for a cross to be evaluated in replicated yield trials. The crosses were derived as described in Part II.

Field Performance Trials

The 28 entries were evaluated for seed yield as described in Part II. In 1990, the two center rows of the check plots were harvested and considered as one

[3] Tables and figures are presented in the Appendix.

observation, and one border was harvested and considered as a separate observation in the same environment. This scheme provided equal numbers of observations for the crosses (1 bulk and 1 line mean) and checks (2) in each 1990 environment. All yields were adjusted to 130 g H_20 kg^{-1}. The purpose of these yield trials was to determine, as accurately as possible, the relative mean seed yield of each cross based on the performances of bulks and lines from the cross.

Historical yield data on the parents (25 total) involved in the 24 crosses were compiled from the UTSB and USDA Soybean Tests - Southern Region conducted in Tennessee (only) from 1982 to 1990. These data consisted of the mean seed yield of a parent from each trial in which the parent occurred. Since some parents had been tested for yield more extensively than others, there was a wide range in the number of records on each parent (Table 2). Three parents, Hutcheson, TN 4-86, and TN 5-85 made up 34% of the total number of historical parent observations; and three parents, Hutcheson, TN 4-86, and TN85-55 were involved in at least six of the 24 total crosses being evaluated. In Part II of this dissertation, it has been demonstrated that these historical parental records could be utilized by BLUP to provide accurate predictions of the ranks of the 24 crosses *a priori*. The intent in this experiment was to determine if their inclusion, along with actual performance data on the 24 crosses, sufficiently enhanced the predictions of the crosses to justify maintaining a large database of parental performance records for future predictions.

Statistical Analyses

Each year/location combination was considered to be a separate environment. Overall least squares mean yields (BLUE) were calculated for both bulks (11 environments) and lines (5 environments) for each cross and check cultivar. All effects were considered fixed. Using rank correlation analysis, it was determined that F_4-F_6 bulk and $F_{5:6}$ line data could be combined to provide the best estimate of the overall mean performance of each cross.

Mean yields of bulks (only) from each cross and check cultivar at each of three locations, KES in eastern Tennessee, HRES in central Tennessee, and AMES in western Tennessee were combined in all combinations of one location in one year (1989 only), two locations in one year (1989 only), and one location across two years (1989-1990) to simulate typical types of data a breeder may possess for estimating yields of F_5-F_6 soybean crosses. These three locations represented the three major climatic and edaphic divisions of Tennessee. There were a total of nine different predictive data sets from which the mean seed yield of each of the 28 entries was estimated using BLUE, and predicted using each of two methods of BLUP. The distinction between an estimate and a prediction is that an estimate provides a value based on any information available up to that point in time; whereas a prediction utilizes all information available to provide a value for some situation in the future. For the purpose of simplicity, the three methods henceforth will be referred to

collectively as methods of prediction. The linear models used to make estimates and predictions are given below:

1) BLUE. The first method was to calculate the least square means (BLUE) for each entry. It is considered the current standard method for *estimation* of means. The additive linear fixed model was:

$$y = \mu + X\beta + \epsilon ,$$

where

y is a vector of measured yield,

μ is the overall mean yield,

X is a design matrix of fixed effects (entry, environment, and blocks within environments combinations),

ß is the vector of unknown effects due to entry, environments, and blocks within environments to be estimated,

ϵ is an error vector with a null mean.

2) BLUP(NP). The second method was to calculate the BLUP as described by Allen *et al.* (1991) for each entry without the inclusion of historical parental data in the predictive data set. This method represented a new approach to mean *prediction* in which related entries (*e.g.* crosses) could contribute to the adjusted yields of one another. The additive linear mixed model was:

$$y = \mu + X\beta + ZU + \epsilon ,$$

where

y is a vector of measured mean yield for each entry in each environment,

μ is the overall mean yield,

X is a design matrix of fixed effects (environments),

ß is the vector of unknown effects due to environments to be estimated,

Z is a design matrix of random effects (entries),

U is the vector of unknown effects due to entries, and

ϵ is an error vector with a null mean.

Because of the limitation of computer resources as the size of the mixed model increased, the y vector consisted of the BLUE least squares mean yield for each entry adjusted for the block within environments effects.

3) BLUP(P). The third method was identical to BLUP(NP) except all historical parental data were included in the predictive set. This method represented a new approach to mean *prediction* in which all available data on the entries *per se*, related entries (*e.g.* crosses), and parents of the entries could be utilized to predict entry means.

Since both BLUP methods required an estimate of heritability for the given trait, variance components were estimated from 1990 $F_{5:6}$ line data using the REML option in SAS® PROC VARCOMP (SAS, 1985) and heritability for yield (0.24, data not shown) was calculated based on these estimates. For each of the BLUP methods, the genetic relationships among entries in the predictive set were incorporated in the

model as estimates of the variance/covariance among random effects (Allen *et al.*, 1991).

When a breeder makes selection decisions, the predicted rank of an individual is generally the major selection criterion. As such, the predicted values (or errors of predictions) of the mean yield are of limited importance, but the variances of the errors of the predictions are usually critical. If the variance of prediction errors is minimized, then the correlation of ranks between the actual and estimated or predicted means should be maximized. Therefore, the three methods were compared for precision of their estimates of predictions by calculating the standard error of the difference ($s_{\bar{d}}$) between the actual and the predicted mean seed yield of each entry for each predictive data set. The $s_{\bar{d}}$ provided a measure of the variance in prediction errors. Rank correlation analysis was used to determine how closely each method ranked entries relative to the actual ranking.

In order to remove any auto-correlation between the actual and the predicted means, an overall mean for each entry was recalculated in each case by excluding data from the predictive set and recomputing what was considered the actual mean from the remaining data. For example, if bulk data from KES 1989 were used as the predictive data set, then KES 1989 bulk data were excluded from the complete data set and a new overall mean was computed using the rest of the data.

CHAPTER 3

Results and Discussion

Actual Yield Performance of Bulks and Lines

Rank correlation between least square mean seed yields across environments for bulks and lines was found to be positive (r_s=0.77) and significant ($P \leq 0.01$). Therefore, data for both bulks and lines were combined to provide the best estimate attainable for each cross and check cultivar (Table 1).

Comparison of Predicted Yields from BLUE, BLUP(NP), and BLUP(P)

BLUP(P) predictions always provided more conservative estimates than either BLUP(NP) or BLUE. As an example, the mean of the predictions made from KES 1989 are 2513, 3329 and 3329 kg ha^{-1}, for BLUP(P), BLUP(NP) and BLUE, respectively (Table 3). This trend was consistent for all predictive data sets (data not

shown). There were two reasons for the conservatism of BLUP(P). First, the overall mean was somewhat lower when historical parental data were included. This decrease was reflected in the predictions made from BLUP(P). Second, BLUP adjusts predictions away from the observed mean and toward the overall mean when few observations are available (Henderson, 1977; Hill and Rosenberger, 1985; Allen *et al.*, 1991). In the cases of predicting means from one location in one year, only one observation was available for each cross while many observations were available for parents of the crosses (Table 2). For BLUP(NP), the shrinkage was relatively equal for each entry because equal numbers of observations were present for each entry. The differences between the BLUP(NP) and BLUE were due both to the genetic relationships that existed among entries and the shrinkage properties of BLUP.

The standard errors (SE) of the predictions were significantly less for both methods of BLUP than for BLUE for all nine predictive data sets (data not shown). However, the magnitude of SE for both methods of BLUP were inconsistent with the magnitude of the data. For example, using KES 1989 as the predictive data sets, CROSS-01 had a predicted mean of 3311, 3231, and 2363 kg ha^{-1} using BLUE, BLUP(NP), and BLUP(P), respectively (Table 3). The SE of those predictions were 147.5, 0.47, and 0.40 kg ha^{-1} using BLUE, BLUP(NP), and BLUP(P), respectively.

To determine why the SE from these two forms of BLUP were about 30 times smaller than those from BLUE, entry means were predicted using another mixed model analysis where:

$$G = I\,(\sigma_g^2 \times h^2)$$

and

$$R = I\,(\sigma_e^2 \times [1 - h^2])\ .$$

This model was comparable to the mixed model used in these analyses with the exception that the genetic relationships were not incorporated in the G matrix. The SE's from this model were approximately 99 kg ha^{-1} for each entry (data not shown) when KES 1989 was used as the predictive data set. Based on these observations, it was concluded that the decreases in SE using BLUP(P) and BLUP(NP) were primarily due to the significant amount of genetic variation explained by the genetic variance/covariance structure in the genetic relationship matrix.

For example, the pedigree of CROSS-04 was TN 4-86/Hutcheson (Table 1). Five other crosses (CROSS-22, CROSS-23, CROSS-10, CROSS-24, and CROSS-18) had Hutcheson as one parent, and six other crosses (CROSS-03, CROSS-11, CROSS-13, CROSS-08, CROSS-01, and CROSS-17) had TN 4-86 as one parent. Since there was a high degree of genetic correlation between CROSS-04 and each of these related crosses, records on these crosses (among others) plus TN 4-86 and Hutcheson increased the *effective* number of observations of CROSS-04. The increase in the effective number of observations of CROSS-04 was reflected in the dramatic

decrease of the SE of its predicted value from BLUP(NP) and BLUP(P) *versus* BLUE (0.45 and 0.40 *versus* 148.3 kg ha^{-1}, respectively, Table 3) when KES 1989 was the predictive set.

The standard deviation (s) among predictions using BLUP(P) and BLUP(NP) were less than those for BLUE (126, 123, and 332 kg ha^{-1}, respectively, Table 3) in KES 1989. Likewise, the standard deviations among the predictions from both methods of BLUP were less than those from BLUE in the other eight predictive data sets (data not shown). This observation is consistent with those of Hill and Rosenberger (1985) and White *et al.* (1986) and indicates the conservative nature of BLUP values when there are small numbers of observations. Under these conditions, predictions are adjusted toward the overall mean and away from the observed mean.

These results indicate that plant breeders should have more confidence in predictions from BLUP than in those from BLUE. If an extremely low or high value is estimated for a genotype from one or two environments using BLUE, prudence would dictate that the breeder should view the estimate cautiously. When BLUP is used to predict the mean of a genotype, the "best estimate" is closer to what the performances of its relatives have already shown to be - thus, the prediction of the mean of the genotype is adjusted toward the mean of its relatives. The magnitude of shrinkage is determined by the degrees of genetic relationships between the genotype and its relatives for which performance data are available. In other words, the "best estimate" is that the genotype is not as good or as bad as it appears to be, based on the established performance of its relatives. If the actual value of the genotype truly

falls significantly beyond the overall mean, accumulating more observations on the genotype *per se* will confirm it, and its subsequent prediction will approach its observed value.

If the SE's of BLUP predictions are more accurate than those of BLUE, their magnitude can have profound implications for genotype selection. When a breeder uses truncated selection based on BLUP values to select a set of superior genotypes from among a larger group, more confidence can be placed in those selections than in selections based on BLUE values. This confidence stems from the increased resolution of BLUP over BLUE. That is, the breeder could distinguish more accurately between two genotypes with similar performance using BLUP since the SE's associated with the predictions from BLUP are much smaller than the SE's associated with the estimates from BLUE for the same two genotypes.

Comparison of Actual *versus* Estimated and Predicted Yields

The $s_{\bar{d}}$ between the actual and estimated or predicted means were calculated for each method for each of the nine data sets (Table 4). In every case except one, the $s_{\bar{d}}$ of both methods of BLUP were less than for BLUE. For the one exception, HRES across two years, there were no differences among $s_{\bar{d}}$ for all three methods (23, 27, and 26 kg ha^{-1} for BLUE, BLUP(NP), and BLUP(P), respectively). The $s_{\bar{d}}$ for BLUP(NP) and BLUP(P) were comparable in every case, indicating that there was no

gain in precision of the predictions using these data sets with the addition of parental data.

Correlation of ranks between the actual and estimated or predicted seed yields for each of the nine predictive data sets showed that both methods of BLUP resulted in consistently higher correlations than BLUE (Fig. 1). The only exception was HRES across two years where the rank correlations were the same for all three methods (0.81, 0.82, and 0.81 for BLUE, BLUP(NP), and BLUP(P), respectively). The anomalous results for HRES across two years may be attributed to a very dry season in 1990. This poor growing season reduced the expression of genetic variation and thereby introduced error for all three prediction methods. There was little difference in rank correlations between BLUP(P) and BLUP(NP).

These analyses indicated that either form of BLUP was superior to BLUE in ranking the 24 crosses in this study for yield performance. Generally, there was no advantage in including historical parental data to supplement predictions made by BLUP because of the high degrees of genetic relationships that existed among the 24 crosses.

CHAPTER 4

Conclusions

Results from these analyses indicate that either form of BLUP is superior to BLUE (least squares estimates) for improving both the precision (*e.g.* SE) and the accuracy (*e.g.* r_s, $s_{\bar{d}}$, and SE) of soybean yield values. Rank correlations of the actual and the predicted means were always higher and SE's of the predictions were always lower using BLUP *versus* BLUE. Based on the parents and crosses used in this study, including parental data in BLUP did not enhance the precision of predictions sufficiently to justify the extra effort involved in maintaining a large database of historical records. However, the genetic material used in this study was highly related (Table 1), and therefore predictions of cross means were greatly affected by observations of other crosses. Under different circumstances, such as predicting mean yield of genetically diverse populations, BLUP(NP) might not have been as effective a predictor of the actual mean as BLUP(P). Therefore, it should be recommended that if historical parental data are available, they should be included because they can only enhance the precision of predictions of progeny.

Literature Cited

Allen, F.L. and H.L. Bhardwaj. 1987. Genetic relationships and selected pedigree diagrams of North American soybean cultivars. Tennessee Agric. Exp. Stn. Bull. 652.

Allen, F.L., D.M. Panter, W.L. Sanders, and R.A. McLean. 1991. Best linear unbiased prediction as a method to enhance breeding for yield in soybeans. Crop Sci. (in review).

Blouin, D.C., A.M. Saxton, and K.L. Koonce. 1989. A completely randomized design with missing cells and repeated measures in time. p. 80-103. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Bridges, Jr., W.C. 1989. Analysis of a plant breeding experiment with heterogeneous variances using mixed model equations. p. 145-154. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Henderson, C.R. 1952. Specific and general combining ability. p. 352-370. *In* J.W. Cowan (ed.) Heterosis. Iowa State University, Ames, IA.

Henderson, C.R. 1963. Selection index and expected genetic advance. p. 141-163. *In* W.D. Hanson and H.F. Robinson (ed.) Statistical Genetics and Plant Breeding. NAS-NRC Pub. No. 982.

Henderson, C.R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-477.

Henderson, C.R. 1975b. Use of all relatives in intraherd prediction of breeding values and producing abilities. J. Dairy Sci. 58:1910-1916.

Henderson, C.R. 1977. Prediction of future records. p. 615-638. *In* E. Pollack, *et al.* (ed.). Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames, IA.

Hill, R.R. and Rosenberger, J.L. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470.

McLean, R.A. 1989. An introduction to general linear models. p. 9-38. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Sanders, W.L. 1989. Choice of models. p. 49-56. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

SAS Institute Inc. 1985. SAS® Users's Guide: Statistics, Version 5 Edition. Cary, NC. SAS Institute Inc. 956 pp.

SAS Institute Inc. 1988. SAS/IML™ Users's Guide for Personal Computers, Version 6 Edition. Cary, NC. SAS Institute Inc. 243 pp.

Stroup, W.W. 1989. Why mixed models? p. 1-8. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

White, T.L., G.R. Hodge, and M.A. Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. p. 99-122. *In* Workshop of the Genetics and Breeding of Southern Forest Trees RIEG, June 25-26, 1986, Univ. of Florida, Gainsville, FL.

Appendix

Table 1. Mean yields and ranks of 24 crosses based on tests of F_4-F_6 bulks (1988-1990), $F_{5:6}$ lines (1990), and combined bulks and lines.

Soybean Cross/ Cultivar	Cross I.D.	Yield			Rank		
		Bulk†	Line‡	Overall§	Bulk	Line	Overall
		------- kg ha^{-1} --------					
Crosses							
Hutcheson/TN82-221	CROSS-22	2645	1882	2589	2	2	1
Hutcheson/TN85-55	CROSS-23	2657	1838	2584	1	3	2
TN85-116/Hutcheson	CROSS-10	2489	1983	2514	4	1	3
D83-3318/Hutcheson	CROSS-24	2548	1828	2506	3	4	4
TN 4-86/Hutcheson	CROSS-04	2372	1806	2378	5	5	5
TN85-3/Hutcheson	CROSS-18	2347	1643	2310	6	8	6
TN84-3/TN85-157	CROSS-06	2310	1651	2287	7	7	7
TN85-102/TN84-21	CROSS-09	2286	1663	2274	8	6	8
TN85-55/TN84-3	CROSS-21	2282	1633	2262	9	9	9
TN85-172/TN85-157	CROSS-16	2254	1574	2224	10	15	10
TN85-162/TN85-172	CROSS-15	2194	1628	2200	13	10	11
TN85-55/TN 5-85	CROSS-19	2237	1502	2190	11	18	12
Stafford/TN 4-86	CROSS-03	2127	1604	2146	17	11	13
TN85-136/TN85-55	CROSS-14	2150	1537	2141	15	17	14
TN85-121/TN85-116	CROSS-12	2115	1596	2135	18	12	15
TN84-7/TN85-55	CROSS-07	2212	1377	2133	12	22	16
TN85-117/TN 4-86	CROSS-11	2104	1579	2123	19	14	17

Table 1 (continued)

Soybean Cross/ Cultivar	Cross I.D.	Yield			Rank		
		Bulk†	Line‡	Overall§	Bulk	Line	Overall
		------- kg ha^{-1} --------					
TN85-55/TN83-26	CROSS-20	2164	1439	2120	14	19	18
TN85-13/TN 4-86	CROSS-13	2077	1564	2099	20	16	19
TN82-221/TN84-7	CROSS-05	2145	1350	2079	16	23	20
TN85-10/TN 4-86	CROSS-08	2071	1432	2054	21	20	21
D83-3349/TN 4-86	CROSS-01	2005	1401	1999	22	21	22
TN85-22/TN 4-86	CROSS-17	1846	1585	1947	24	13	23
D83-3349/V81-141	CROSS-02	1966	1311	1944	23	24	24
Check Cultivars							
Hutcheson				2767			
TN 5-85				2266			
Essex				2262			
TN 4-86				2012			

† Least squares means of three replications at one location in 1988 and five locations in 1989 and 1990 (11 environments).
‡ Least squares means of 40 lines (eight lines at each of five locations (environments)) in 1990.
§ Least squares means of bulks at 11 environments and lines at five environments (1988-1990).

Table 2. Yield and background information of the 25 parents used in the 24 crosses in this study.

Parent	Pedigree	Historical Yield Records	Parent in Crosses	Yield†	SE†
		----- no. -----		-- kg ha^{-1} --	
D83-3318	Peking/Centennial// Forrest///Bedford	4	1	2078	244
D83-3349	Peking/Centennial// Forrest///Bedford	6	2	2142	211
Hutcheson	V68-1034/Essex	32	6	2750	95
Stafford	V66-318/V68-2331	25	1	2547	105
TN4-86	Beford/Crawford	24	7	2423	107
TN5-85	D68-127/Essex	31	1	2728	92
TN82-221	J74-45/Ts72-824	8	2	2582	214
TN83-26	J74-40/K1017	13	1	2694	138
TN84-21	Bedford/RA401	6	1	2630	195
TN84-3	J74-45/V76-482	6	2	2317	195
TN84-7	J74-45/V76-482	6	2	2154	196
TN85-10	D77-18/Fayette	13	1	2502	137
TN85-102	Jeff/Nathan	8	1	2565	170
TN85-116	Essex/D72-8927	2	2	2014	327
TN85-117	Essex/D72-8927	13	1	2523	137
TN85-121	D72-8927/Epps	7	1	2366	182
TN85-13	Essex/D72-8927	4	1	2515	235
TN85-136	AS5618/D72-8927	4	1	2633	235
TN85-157	D72-8927/TN80-83	16	2	2634	129
TN85-162	Fayette/TN80-57	4	1	3003	235
TN85-172	TN80-70/D72-8927	4	1	2019	237
TN85-22	AS5474/D72-8927	4	1	2230	235
TN85-3	D77-18/PI416.772	4	1	1743	235
TN85-55	TN77-46/Fayette	9	6	2688	161
V81-141	Essex/V71-793	2	1	2057	342

† Least squares means and SE based on historical yield records.

Table 3. Actual and estimated or predicted means from one environment (KES 1989) of four check cultivars and 24 soybean crosses using three methods (BLUE, BLUP(NP), and BLUP(P)).

	Yield + SE							
			Estimated‡		Predicted‡			
Cultivar	Actual†		BLUE		BLUP(NP)		BLUP(P)	
	---------- kg ha^{-1} ----------							
Crosses								
CROSS-01	1875	± 70.7	3311	± 147.5	3231	± 0.47	2363	± 0.40
CROSS-02	1825	± 70.7	3179	± 147.3	3317	± 0.48	2348	± 0.43
CROSS-03	2057	± 70.7	2940	± 148.3	3169	± 0.49	2435	± 0.40
CROSS-04	2277	± 70.7	3343	± 148.3	3345	± 0.45	2572	± 0.40
CROSS-05	1946	± 70.7	3537	± 148.3	3388	± 0.50	2463	± 0.42
CROSS-06	2191	± 70.7	3175	± 148.3	3316	± 0.49	2459	± 0.41
CROSS-07	2009	± 70.7	3448	± 148.3	3379	± 0.47	2506	± 0.42
CROSS-08	1953	± 70.7	3025	± 148.3	3182	± 0.48	2422	± 0.41
CROSS-09	2146	± 70.7	3652	± 147.9	3385	± 0.51	2580	± 0.42
CROSS-10	2420	± 70.7	3368	± 147.9	3428	± 0.48	2557	± 0.42
CROSS-11	2031	± 70.7	2956	± 147.5	3192	± 0.46	2412	± 0.40
CROSS-12	2030	± 70.7	3167	± 147.9	3276	± 0.51	2369	± 0.43
CROSS-13	2008	± 70.7	2918	± 147.5	3185	± 0.46	2399	± 0.42
CROSS-14	2030	± 70.7	3251	± 148.3	3292	± 0.50	2568	± 0.43
CROSS-15	2101	± 70.7	3140	± 148.3	3252	± 0.48	2502	± 0.42
CROSS-16	2114	± 70.7	3327	± 147.9	3293	± 0.51	2432	± 0.42
CROSS-17	1876	± 70.7	2464	± 148.3	3068	± 0.48	2258	± 0.42
CROSS-18	2196	± 70.7	3473	± 148.3	3440	± 0.49	2455	± 0.42

Table 3 (continued)

Cultivar	Actual†		Yield + SE: Estimated‡ BLUE		Predicted‡ BLUP(NP)		Predicted‡ BLUP(P)	
	---------- kg ha^{-1} ----------							
CROSS-19	2071	± 70.7	3427	± 148.3	3383	± 0.46	2662	± 0.40
CROSS-20	2024	± 70.7	3013	± 147.5	3259	± 0.49	2581	± 0.41
CROSS-21	2132	± 70.7	3658	± 147.5	3404	± 0.47	2559	± 0.42
CROSS-22	2487	± 70.7	3578	± 147.5	3463	± 0.47	2655	± 0.40
CROSS-23	2453	± 70.7	3994	± 147.5	3521	± 0.46	2766	± 0.41
CROSS-24	2387	± 70.7	3737	± 148.3	3473	± 0.46	2565	± 0.40
Cultivars								
Essex	2125	± 70.7	3759	± 147.5	3471	± 0.44	2608	± 0.36
Hutcheson	2660	± 73.4	3825	± 147.5	3552	± 0.43	2757	± 0.20
TN4-86	1903	± 70.7	3100	± 148.3	3138	± 0.40	2415	± 0.21
TN5-85	2150	± 73.4	3448	± 147.9	3410	± 0.47	2692	± 0.19
Overall $\bar{X}$	2124		3329		3329		2513	
s			332		123		126	

† Least squares means based on F_4-F_6 bulks at one location in 1988, five locations in 1989, four locations in 1990, and $F_{5:6}$ lines at five locations in 1990 (total of 11 environments).
‡ Least squares means based on F_5 bulks at KES in 1989.

Table 4. Standard errors of the differences ($s_{\bar{d}}$) between actual and estimated or predicted mean seed yield of 28 entries (four check cultivars and 24 crosses) for three different methods (BLUE, BLUP(NP), and BLUP(P)) using combinations of one location in one year, one location across two years, and two locations in one year as predictive data sets.

Predictive Data Set	$s_{\bar{d}}$		
	BLUE	BLUP(NP)	BLUP(P)
	------- kg ha^{-1} -------		
One Location in 1989			
KES	48	26	26
HRES	37	27	26
AMES	70	23	23
One Location Across 2 Years (1989-1990)			
KES	50	23	18
HRES	23	27	26
AMES	55	24	24
Two Locations in 1989			
KES + AMES	44	22	21
KES + HRES	40	25	25
HRES + AMES	41	22	22

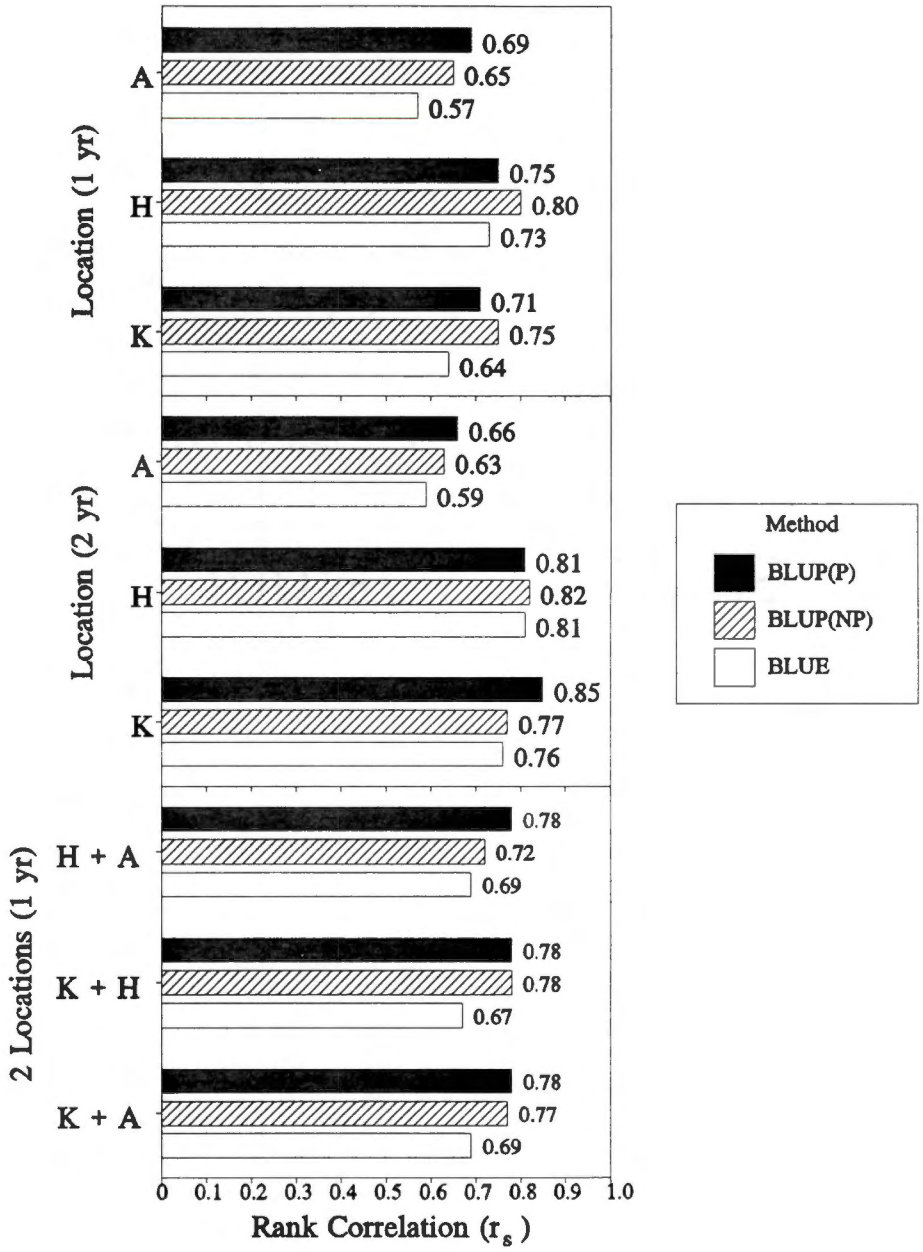


Fig. 1. Rank correlations between actual and estimated (BLUE) or predicted (BLUP(P) and BLUP(NP)) means of the 24 crosses using different combinations of three locations (A=AMES, H=HRES, and K=KES) across two years.

PART V

A Computing Strategy for Utilizing Mixed Linear Models and BLUP in Plant Breeding Applications

CHAPTER 1

Introduction

When breeding for yield, plant breeders historically have faced the problem of parental selection in order to obtain hybrid populations with high expected mean yield performance and significant genetic variation. Identification of cross combinations which meet the above criteria result in higher probabilities of selecting progenies with sufficient genetic superiority to warrant potential variety release. Typically, predicting the mean of cross combinations is made by calculating midparent values (MPV), or the mean of parental means, based on observed records of potential parents. In order to have the most confidence in MPV as a predictor of future progeny, it is important to obtain the best unbiased estimate of a parental mean as possible.

Best Linear Unbiased Estimates (BLUE)

Achieving the best *estimate* of a single trait mean often is accomplished by employing some form of an additive linear model. With such a model, effects which contribute to observed values can be estimated and "adjusted out", and an "unbiased" estimate of the genotypic effect can be calculated. A classical method of obtaining the adjusted estimate is by considering all effects fixed and combining data from field tests across locations and/or years. This approach results in ordinary or weighted least squares solutions (White *et al.*, 1986). The effects due to locations, years, and blocks within location/year are "adjusted out" and the best linear unbiased estimate (BLUE) of the genotypic effect is isolated. However, in cases where only unbalanced data (individuals are not equally represented across years, locations, and blocks) are available, the fixed effects approach can lead to imprecise genotypic estimates resulting in a tendency to select genotypes which are included in fewer tests (Henderson, 1973; White *et al.*, 1986). If the fixed effects approach is used to estimate parental breeding values and no data are available for some of the parents of interest, no estimates of their breeding values are possible.

Best Linear Unbiased Predictions (BLUP)

Henderson (1975a; 1977) described the use of a mixed linear model to calculate the best linear unbiased *predictions* (BLUP) of breeding values of potential parents based on observed records and known (or estimated) variance/covariance structure among random effects. The distinction between this method of *prediction* and the BLUE method of *estimation* is subtle, but important. An estimate uses all information available to provide a value at that point in time; whereas a prediction uses all information available to provide a value which applies to some situation in the future.

Henderson showed that the predictions of the random effects and estimates of the fixed effects (u and b, respectively) could be obtained by any solution of the following equations:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} * \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

where X and Z design matrices of fixed and random effects, respectively, and R and G are the variance/covariances of error and the random effects, respectively, and Y is the vector of observed responses. The prime (') symbol represents the transpose of a matrix and the superscript, -1, represents the inverse of a matrix. For genetic applications, Henderson's methods utilize the genetic relationships among individuals as the variance/covariance among genetic effects and assume that the only cause of

correlation between observations on different individuals is additive genetic variance (1975b). Other effects, such as years and locations, can be considered fixed.

By utilizing the genetic variance/covariance among individuals, related individuals contribute to predicted values of one another. The obvious advantage of this approach for predicting genotypic means is that when no data are available on some genotypes, their values can still be predicted from the observed values of their relatives. In Part II of this dissertation, the methodology has been shown to be effective in choosing parental cross combinations when little or no data are available on potential parents. In such situations, breeding values can be predicted for any parent where information on the parent *per se* or its relatives is available; and predictions of crosses can be made where more classical methods fail. This approach can be utilized also for evaluating new genotypes in the early stages of performance testing when little information on the genotype is available. Data from related individuals (*e.g.* parents, sibs) in past yield trials can be used to supplement information in early performance tests to improve the predictions of new genotypes.

Henderson applied this methodology for predicting breeding values of dairy cattle sires based on the milk yield of their dams and daughters. He demonstrated later (1977) how a narrow sense estimate of heritability (h^2) of the trait of interest

also could be incorporated into the model. Since an estimate of error is included in h^2, the error matrix (R) can be canceled from the mixed models equations. With the availability of an estimate of heritability, the equations become:

$$\begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix} * \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\frac{1-h^2}{h^2} \end{bmatrix}^-$$

where A is the matrix of genetic relationships among individuals and h^2 is the heritability of the trait. The superscript minus sign represents the generalized inverse of a matrix.

Another property of BLUP which is advantageous to plant breeders is the shrinkage of predictions of individuals for which there is little information. This shrinkage tends to "adjust" predictions back toward the overall mean when there are small numbers of observations of a particular individual, and "adjust" predictions closer to the observed individual mean as the number of observations increases. The concept of shrinkage has been described in combining unbalanced data from alfalfa (*Medicago sativa* L.)(Hill and Rosenberger, 1985) and soybean (*Glycine max* (L.) Merr.) yield trials (Sanders, 1989), and balanced data from cucumber (*Cucumis sativus* L.) yield trials across years and/or locations (Bridges, 1989).

There is very limited information available on the use of BLUP for plant breeding applications where the genetic relationships among individuals are utilized as the variance/covariance structure among random genetic effects. Allen and

Bhardwaj (1987) used BLUP to predict the response of soybean cultivars and ancestral lines to simulated acid rain. Allen *et al.* (1991) later demonstrated, with simulated data, the potential effectiveness of BLUP for predicting yield performance of future soybean crosses from performance of their relatives. These papers represent the only published use of BLUP under such conditions since Henderson's (1977) work.

Mixed models analyses might not have been utilized widely for two primary reasons. First, until recently there was very little published research explaining mixed models methodology and its potential usefulness in agricultural research. The papers mentioned above, among others, should provide the necessary background for more researchers to utilize this methodology. Second, computer programs which could perform mixed models analyses were unavailable. In order to solve the mixed models equations, programs had to be written in matrix languages such as SAS/IML™ (SAS, 1985). Blouin *et al.* (1989) described a computer program, called GLMM, which represents the first publicly available software capable of performing mixed models analyses without the necessity of writing computer code. Unfortunately, GLMM assumes no covariance among random effects (*e.g.* individuals) and is therefore ineffective for genetic applications.

Some of the recent research in our program has led us to believe that BLUP will have significant impact in plant breeding applications. In order for plant breeders to employ this methodology, it is necessary to provide a detailed description of the computing algorithms involved in genetic analyses.

The objectives of this report are to i) demonstrate a computing strategy for using mixed models analysis in genetic applications, ii) describe and provide examples of the components of a mixed models analysis for predicting seed yield of a future soybean cross, and iii) provide all the computer source code used for making the predictions.

All of the computer programs presented here were written in SAS/IML™ and have been verified to run under SAS® Version 6.04 for Personal Computers, and SAS® for OS/2™ Version 6.06 on an IBM™ Model 80 personal computer, and under SAS® Version 6.06 on an IBM™ 3090 mainframe computer running under the Conversational Monitor System (CMS) operating system. We hope that the details of this report will provide the necessary base of information in both theory and application to encourage plant breeders to utilize this potentially valuable methodology.

CHAPTER 2

Overview

The basic computing strategy we employ to perform mixed models analyses for plant breeding applications has three major steps (Fig. A-1).

Step 1: uses a computing algorithm described by Henderson (1975c) to compute recursively a precusor matrix (L matrix) to the genetic relationship matrix (A matrix) among individuals. This step requires an input data set consisting of a list of pedigrees of the individuals of interest. In the case of our soybean breeding program, the pedigrees of all individuals of interest have been traced back to their original ancestral lines (Allen and Bhardwaj, 1987) and the data set can include more than 900 different pedigrees.

Step 2: generates the A matrix from the L matrix using a data set consisting of a list of potential individuals for which a prediction of mean yield may be desired. This step is required to limit the size of the A matrix

to conform to the computing resources available. For example, if we are interested in predicting only the mean of 100 individuals, the second step generates an A matrix with the dimensions of 100 x 100 corresponding to the genetic variances (relationship of an individual with itself) and covariances (relationships between individuals) among the 100 individuals listed in the data set. If all individuals from the pedigree data set in first step are to be predicted, this step merely creates the A matrix for all individuals.

Step 3: combines yield data collected on the individuals of interest with their corresponding A matrix to calculate the BLUP and standard error of the prediction (SE) for each individual. The yield data set consists of the mean of an individual at each environment in which it was evaluated. In this step, we fit the additive mixed linear model:

$$y = \mu + X\beta + ZU + \epsilon ,$$

where

y is a vector of measured mean yield for each entry in each environment,

μ is the overall mean yield,

X is a design matrix of fixed environments,

ß is the vector of unknown effects due to environments to be estimated,

Z is a design matrix of random individuals,

U is the vector of unknown effects due to individuals, and

ϵ is an error vector with a null mean.

The mixed models equations yield the solution vector:

$$Solution = \begin{bmatrix} \mu \\ b_{i=1} \\ . \\ . \\ b_{i=p} \\ u_{j=1} \\ . \\ . \\ u_{j=q} \end{bmatrix}$$

where

μ = overall mean yield,

b_i = the effect due to i^{th} environment,

u_j = the effect due to j^{th} individual,

p = total number of environments, and

q = total number of individuals.

The BLUP of the j^{th} individual is defined as the overall mean plus the effect due to individual j plus the average effect of all environments. It is calculated as:

$$BLUP_j = \mu + u_j + \frac{\sum b_i}{p} \quad .$$

The standard error of a prediction is calculated as described by McLean (1989).

The source code for each program is presented in Appendix B and is written for SAS® for Personal Computers Version 6.04. Minor modifications to the source code will be required for specific operating systems.

Computer Hardware/Software System Specifications

For the purposes of this report, we will assume that the reader will be using SAS® for Personal Computers Version 6 or greater. We will assume also that a directory resides on the reader's C disk with the name BLUP. All SAS/IML™ programs should be stored in C:\BLUP and will be referred to as *program*.SAS where *program* is the name of the SAS/IML™ program. All input data sets are permanent SAS® data sets and must be created using a LIBNAME and DATA step in SAS® (SAS, 1988); all output SAS® data sets will be stored as permanent SAS®

data sets in the C:\BLUP directory. The libname SAVE will be used to point to C:\BLUP in all SAS/IML™ programs presented here. SAS® data sets will be referred to as SAVE.*name*, where *name* is the name of the permanent SAS® data set.

The following sections describe each of the three steps of our computing strategy in detail. A hypothetical example is provided to illustrate the contents of each data set and the output generated by each SAS/IML™ program.

CHAPTER 3

Example Problem

The following hypothetical example will be used to demonstrate the input required and output generated for each of the three steps in computing BLUP. Assume that there are five individuals with the pedigree relationships given in Fig. A-2. Individuals A, B, and D are considered unrelated, C was derived from a cross between A and B, and E was derived from a cross of C and D. Furthermore, it is assumed that individuals C and E are the only ones among this group for which we have yield performance data. The goal is to predict the mean yield of a cross between A and D which we will name F (Fig. A-2). Using BLUP, the breeding values of A, B, and D, and the mean performance of future progeny F (*i.e.* cross mean) will be predicted.

Step 1 - Generating the L Matrix. Henderson (1975c) described a lower triangular matrix, called L, as having properties such that: 1) LL' = A, and 2) it could be generated directly from a list of pedigrees. Our SAS/IML™ program, MAKEL.SAS (Appendix B-1), creates the L matrix and stores it as a vector which

will be utilized later. In order to generate the L matrix, it is necessary to create the SAS® data set named SAVE.PEDLIST, which has a list of the pedigrees of the individuals involved in our example problem (Table A-1). The data set SAVE.PEDLIST has four variables, FEMALE, MALE, OFF, and INDEX. The first three variables are character variables with the names of the female and male parents and the offspring, respectively. The last variable, INDEX, is a numeric variable identifying the offspring with a unique index number. *Because the MAKEL.SAS program generates a dummy record with an index value of 1 to initialize the SAVE.L data set, the index value for the first record must be 2 and all subsequent records should follow in sequence. This data set must be ordered such that pedigrees of parents precede those of their progeny.*

To speed computational time, the MAKEL.SAS program requires that the list of pedigrees be read into the program in a numeric format rather than the character format in the SAVE.PEDLIST data set. The program MAKEL.SAS calls a second program, INDEX.SAS (Appendix B-2), which converts the character pedigrees into numeric pedigrees. The INDEX.SAS program uses the index number from the SAVE.PEDLIST data set to recode all parents and offspring to their corresponding index numbers. The output data set from the INDEX.SAS program is then read back into the MAKEL.SAS program to generate L.

The output from the MAKEL.SAS program is written directly to a SAS® data set named SAVE.L. Table A-2 shows the data set SAVE.L for our example with seven individuals (six individuals from SAVE.PEDLIST and one dummy individual).

There is a total of 28 observations which correspond to the lower triangular of the 7 x 7 L matrix. Observation 1 corresponds to the dummy individual, observation 2 corresponds to individual A with the dummy, observation 3 corresponds to A with A, *etc.* The SAVE.L data set is converted later into the L matrix (Fig. A-3). The first row and column of the matrix correspond to the dummy individual created by the MAKEL.SAS program, the second row and column correspond to individual A, the third row and column to individual B, and so on.

Step 2 - Generating the A Matrix. The second step in the analysis is to create the A matrix from L using the SAS/IML™ program BIGA.SAS. The BIGA.SAS program (Appendix B-3) requires an input SAS® data set with one character variable, PARENT. For this example, the data set has been named SAVE.PARENTS. This data set is a listing of the individuals for which rows and columns will be extracted from L (*e.g.* a potential set of parents for crossing). In the case of our example, we are interested in all six individuals and will therefore include the names of all the individuals in the SAVE.PARENTS data set (Table A-3). If many pedigrees are listed in the SAVE.PEDLIST data set in Step 1 above, and only a small subset of individuals will be used in future analyses, the SAVE.PARENTS data set would have only the names of the individuals in the subset. You may include as many individuals as you wish in the SAVE.PARENTS data set but remember that storage of a large A matrix requires a large amount of

disk space. In practice, we include all individuals in the data set SAVE.PARENTS which might be evaluated in later analyses.

The output from the BIGA.SAS program is a SAS® data set, SAVE.BIGA, which includes the variables ENTRY, INDEX, and COL1 to COLn, where n is the number of columns in the A matrix. ENTRY is the character name of the individual, INDEX is the index number of the individual, and COL1 to COLn are the corresponding columns of the A matrix. The SAVE.BIGA data set for our example is presented in Table A-4.

In some large genetic applications, the A matrix may be so large that the computer resources required for its inversion may be unavailable. Henderson (1977) described a computing algorithm which can create A^{-1} as the pedigree data are read. We have written a SAS/IML program which uses Henderson's algorithm to create A^{-1} from the SAVE.PEDLIST data set. The source code for this program is not presented in this report.

Step 3 - Calculating BLUP from Yield Data. The SAS/IML™ program PREDICT.SAS (Appendix B-4) computes the BLUP and SE for each individual listed in the input SAS® data set SAVE.YIELD. The data set SAVE.YIELD must contain three variables, ENTRY, ENV, and YIELD. ENTRY is the character name of any individual, ENV is a numeric variable corresponding to the environment in which the ENTRY was evaluated, and YIELD is the measured response. The program PREDICT.SAS reads the SAVE.YIELD data set to determine which

individuals have corresponding values in the A matrix stored in the SAVE.BIGA data set. If an individual is found in the SAVE.YIELD data set which does not have a corresponding record in the SAVE.BIGA data set, records for that individual are discarded from the analysis. *If the yield of an individual is to be predicted but there are no observed yield data for the individual, it must still be included in the SAVE.YIELD data set. When inputting data into the SAVE.YIELD data set a record for an individual which has no observed values for YIELD, be sure to put a value for ENV that is the same as one of the ENV values of an individual for which there is actual yield data. Otherwise, the record will be dropped from the analysis because the PREDICT.SAS program will not estimate the effect for an environment when no data are available in that environment.* The PREDICT.SAS program creates a temporary A matrix (*i.e.* a subset of the complete A matrix) including only the individuals in the SAVE.YIELD data set and then computes the BLUP and SE for each individual. The results from the analysis are printed through SAS® PROC PRINT.

The information contained in the data set SAVE.YIELD for our example is given in Table A-5. Notice that only individuals C and E have observed values for YIELD, and individuals A, B, D, and F are all included but have missing values for YIELD. Notice also that the value for ENV is 1 for A, B, D, and F, which is equal to the value for ENV for some of the records on C and E. For both individuals C and E, there is one record (*e.g.* the mean of *n* replications) in each of four environments.

The BLUP and SE for each of the six individuals are given in Table A-6. Notice that although there were no observations for A, B, D, and F, their means can still be predicted using BLUP. The predicted values of A and B are based on their observed progeny C and grand-progeny E; the predicted value for D is based on its observed progeny E. The predicted value for F is the mean of the predicted values (*i.e.* breeding values) for its parents, A and D.

The SE's reflect the number of observations for each individual. Notice that C and E have SE of 0.44 units, about half as much as the other four individuals. Even though A and B had no observations, their SE's are less than for D because two of their progeny, C and E had observations, while D had only one progeny (E) for which observations were available. Allen *et al.* (1991) give a more detailed description of the effects of progeny on the predictions of parents.

Updating the L Matrix

When the genetic relationships among only a few individuals are required, a satisfactory approach to updating the SAVE.L data set is to update the SAVE.PEDLIST data set and rerun the programs described in this report. As the number of individuals increases, the time required to recompute the L matrix when new individuals are added becomes restrictive. A more efficient approach would be to add new individuals in such a way that data are appended directly to the existing vector of the L matrix stored in the SAVE.L data set.

The program MAKEL.SAS was designed to be able to update the data set SAVE.L as pedigrees on new individuals become available. The source code for the program (named UPDATEL.SAS) is not presented in this report. However, the MAKEL.SAS program can be modified to function as an UPDATEL.SAS program. The following is a general outline of the steps required to create an UPDATEL.SAS program from the MAKEL.SAS program.

1) Delete lines 10 through 25 of the MAKEL.SAS program (Appendix B-1). These commands create the data sets SAVE.L and SAVE.PEDIGREE. The data set SAVE.PEDIGREE is a numeric list of the parents and offspring used to create the SAVE.L data set. Both of these data sets need to be retained in order to update L.

2) Create a new data set comparable to the SAVE.PEDLIST data set with the pedigrees of the individuals to be added to the SAVE.L program. Be sure that each parent in the new data set appears as an offspring before it appears as a parent. If the parent was previously defined in the SAVE.L and SAVE.PEDIGREE data sets, it can be included in the new data set without reestablishing its pedigree. We will assume the name of the new data set is SAVE.NEWPED.

3) Write a program similar to INDEX.SAS that can index the data set SAVE.NEWPED. The program should read the index numbers in the SAVE.PEDLIST data set and apply them to parents in the SAVE.NEWPED data set where applicable. Output from this program should be a temporary SAS® data set named WORK.A and should include four numeric variables: FEMALE, MALE, OFF, and INDEX representing the numeric index values of the female and male parents, offspring for both variables OFF and INDEX (*i.e.* OFF=INDEX), respectively. Supply the WORK.A data set to the UPDATEL.SAS program, and the data sets SAVE.L and SAVE.PEDIGREE will be updated.

4) Concatenate the SAVE.NEWPED data set onto the SAVE.PEDLIST data set so that the BIGA.SAS program will function properly when a new A matrix is created.

CHAPTER 4

Summary

Using mixed linear models to calculate BLUP should be a very useful tool for predicting means of future crosses in self-pollinated crops. This methodology should improve also the predictions of the performance of new genotypes in the early stages of yield testing. The properties of this methodology make it very applicable to real-world situations faced by plant breeders.

When there are no data on potential parents, predictions of their breeding values can be made based on observed values of their relatives, and the predicted breeding values of parents are adjusted as a function of the amount of information available for them, both *per se* and via relatives (Henderson, 1977; Allen *et al.*, 1991). These properties help provide least biased predictions of breeding values of parents in situations where data are limiting. Since breeding values can be predicted for any parent where information on the parent *per se* or relatives of the parent is available, predictions of crosses can be made using BLUP in situations where more classical methods fail. When used to predict the values of genotypes in the early stages of performance testing, BLUP should improve the precision of predictions by

utilizing any information available on relatives of the genotypes of interest. The potential benefit of BLUP in these situations is that genotype evaluations may be conducted in fewer environments and still provide accurate relative rankings.

A major factor which has limited the use of BLUP has been the lack of computer software capable of performing the analyses for genetic applications. This report demonstrates SAS/IML™ computer programs which can be used either "as is" or modified for specific experiments for calculating BLUP in plant breeding applications. In cases where more complicated models may be required, the program PREDICT.SAS can be modified to accommodate specific applications. The programs MAKEL.SAS and BIGA.SAS may also have broader applications in situations where the genetic relationships among a group of individuals are desired. Using the SAS/IML™ programs presented here, we have begun to evaluate the utility of BLUP for predicting seed yield in soybeans.

Literature Cited

Allen, F.L. and H.L. Bhardwaj. 1987. Genetic relationships and selected pedigree diagrams of North American soybean cultivars. Tenn. Agric. Exp. Stn. Bull. 652.

Allen, F.L., D.M. Panter, W.L. Sanders, and R.A. McLean. 1991. Best linear unbiased prediction as a method to enhance breeding for yield in soybeans. Crop Sci. (in review).

Blouin, D.C., A.M. Saxton, and K.L. Koonce. 1989. A completely randomized design with missing cells and repeated measures in time. p. 80-103. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Bridges, Jr., W.C. 1989. Analysis of a plant breeding experiment with heterogeneous variances using mixed model equations. p. 145-154. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Henderson, C.R. 1973. Sire evaluation and genetic trends. p. 10-41. *In* Animal Breeding and Genetics Symposium in Honor of J.L. Lush, ASAS and ADSA, Champaign, IL.

Henderson, C.R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-477.

Henderson, C.R. 1975b. Use of all relatives in intraherd prediction of breeding values and producing abilities. J. Dairy Sci. 58:1910-1916.

Henderson, C.R. 1975c. Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci. 58:1727-1730.

Henderson, C.R. 1977. Prediction of future records. p. 615-638. *In* E. Pollack, *et al.* (ed.). Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames, IA.

Hill, R.R. and Rosenberger, J.L. 1985. Methods for combining data from germplasm evaluation trials. Crop Sci. 25:467-470.

McLean, R.A. 1989. An introduction to general linear models. p. 9-38. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

Sanders, W.L. 1989. Choice of models. p. 49-56. *In* Applications of mixed models in agricultural and related disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge.

SAS Institute Inc. 1985. SAS/IML™ Users's Guide for Personal Computers, Version 6 Edition. Cary, NC. SAS Institute Inc. 243 pp.

SAS Institute Inc. 1988. SAS® Language Guide for Personal Computers, Release 6.03 Edition. Cary, NC: SAS Institute Inc. 558 pp.

White, T.L., G.R. Hodge, and M.A. Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. p. 99-122. *In* Workshop of the Genetics and Breeding of Southern Forest Trees RIEG, June 25-26, 1986, Univ. of Florida, Gainsville, FL.

Appendix A

Table A-1. Pedigrees (FEMALE and MALE) of individuals (OFF) and the INDEX value assigned to the input SAS® data set SAVE.PEDLIST for the example problem. Pedigrees of individuals must be preceded by the pedigrees of their parents.

	Value			
Observation	FEMALE	MALE	OFF	INDEX
1	.†	.	A	2‡
2	.	.	B	3
3	A	B	C	4
4	.	.	D	5
5	C	D	E	6
6	A	D	F	7

† Indicates a missing value.
‡ Index numbers must start at 2.

Table A-2. Output data set SAVE.L for the example problem.

	Value
Observation	X†
1	1.00000
2	0.00000
3	1.00000
4	0.00000
5	0.00000
6	1.00000
7	0.00000
8	0.50000
9	0.50000
10	0.70711
11	0.00000
12	0.00000
13	0.00000
14	0.00000
15	1.00000
16	0.00000
17	0.25000
18	0.25000
19	0.35355
20	0.50000
21	0.70711
22	0.00000
23	0.50000
24	0.00000
25	0.00000
26	0.50000

Table A-2 (continued)

Observation	Value X†
27	0.00000
28	0.70711

† Values of X correspond to the lower triangular of the 7 x 7 L matrix (See Fig. A-3).

Table A-3. Value of the one variable, PARENT, in the input SAS® data set, SAVE.PARENTS, for the example problem.

	Value
Observation	PARENT
1	A
2	B
3	C
4	D
5	E
6	F

Table A-4. Values of the eight variables ENTRY, INDEX, and COL1 through COL6 contained in the output SAS® data set SAVE.BIGA for the example problem.

	Value							
Observation	ENTRY	INDEX	COL1†	COL2	COL3	COL4	COL5	COL6
1	A	2	1.000	0.000	0.500	0.000	0.250	0.500
2	B	3	0.000	1.000	0.500	0.000	0.250	0.000
3	C	4	0.500	0.500	1.000	0.000	0.500	0.250
4	D	5	0.000	0.000	0.000	1.000	0.500	0.500
5	E	6	0.250	0.250	0.500	0.500	1.000	0.375
6	F	7	0.500	0.000	0.250	0.500	0.375	1.000

† Variables COL1-COL6 correspond to entries A-F, respectively.

Table A-5. Values of the three variables, ENTRY, ENV, and YIELD contained in the input SAS® data set SAVE.YIELD for the example problem.

	Value		
Observation	ENTRY	ENV	YIELD†
1	C‡	1	40
2	C	2	50
3	C	3	45
4	C	4	45
5	E§	1	30
6	E	2	35
7	E	3	40
8	E	4	30
9	A	1	.¶
10	B	1	.
11	D	1	.
12	F	1	.

† Units are generic.
‡ Overall mean of C=45 units; SE=1.69 units.
§ Overall mean of E=34 units; SE=1.69 units.
¶Indicates a missing value.

Table A-6. BLUP and standard errors of predictions (SE) of mean yield for six hypothetical individuals based on performance data from individuals C and E in four environments.

Individual	BLUP	SE
	------ units† ------	
A	40.98	0.88
B	40.98	0.88
C	42.59	0.44
D	36.16	0.93
E	36.16	0.44
F	38.57	0.93

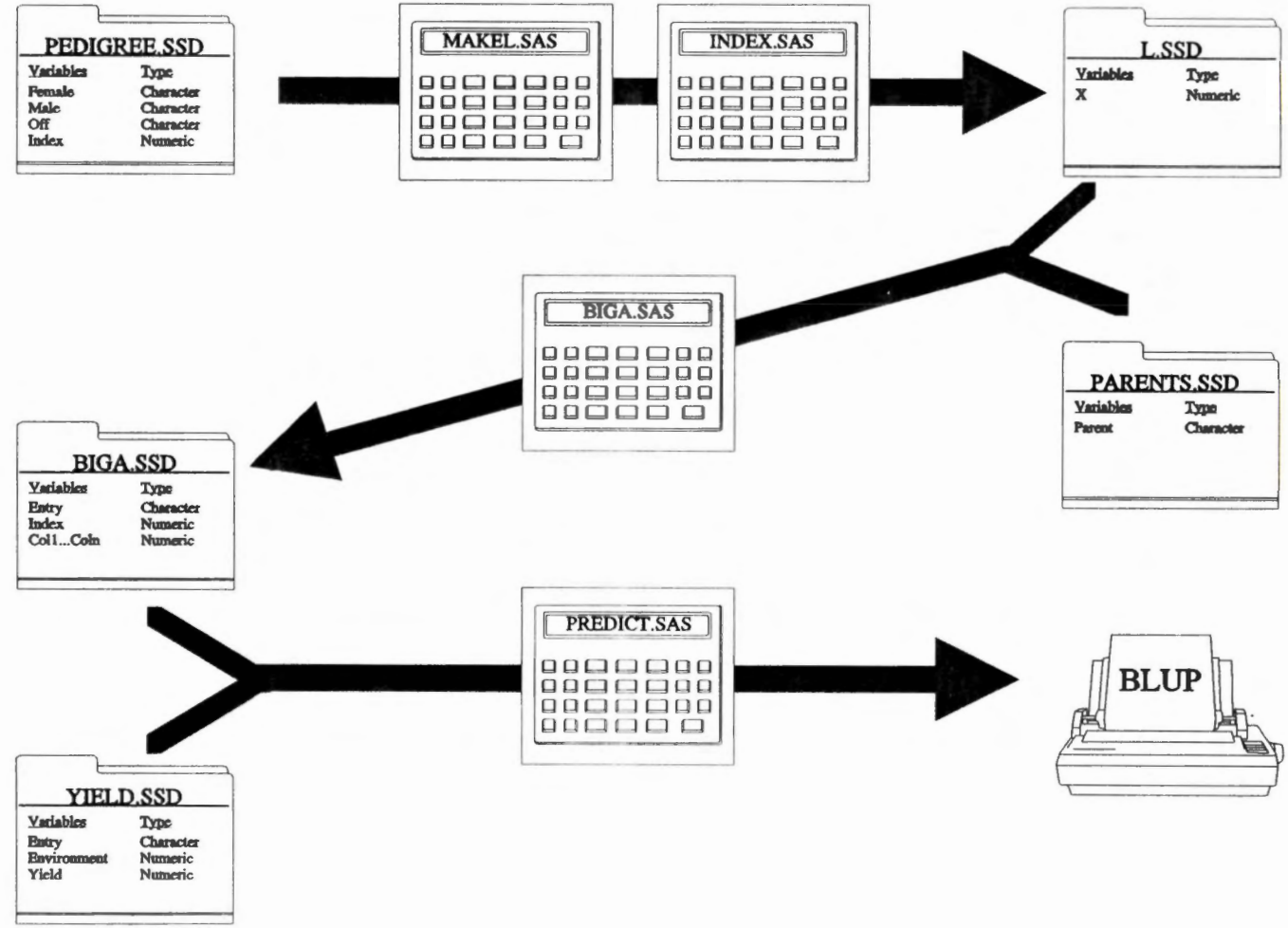
† Generic.

Fig. A-1. Flow diagram of the relationships among SAS/IML™ programs and SAS® data sets used in predicting genotypic means from BLUP. SSD and SAS extensions refer to permanent data sets and programs, respectively.

SAS Data Sets

SAS/IML Programs

SAS Data Sets



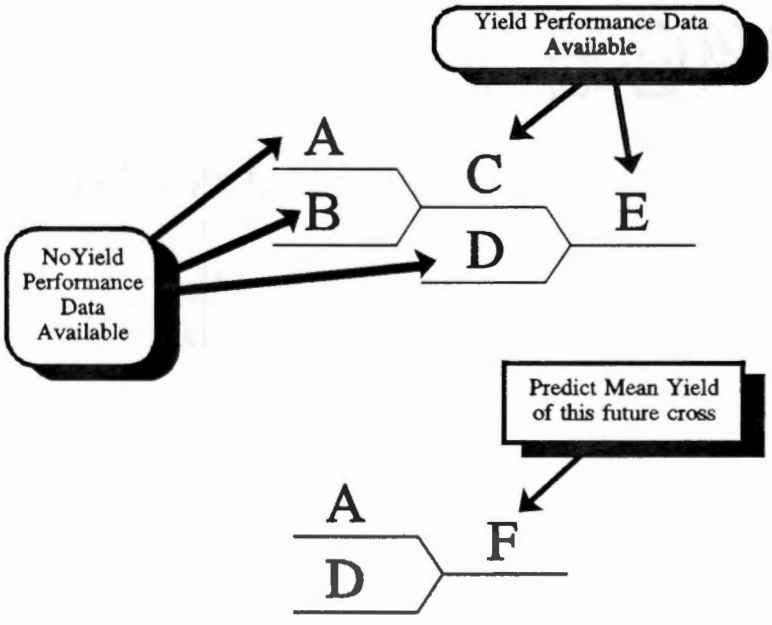


Fig. A-2. Hypothetical pedigree relationships among five individuals (A-E) and pedigree of the future cross F for the example problem.

	Dummy	A	B	C	D	E	F
Dummy	1.00000						
A	0.00000	1.00000					
B	0.00000	0.00000	1.00000				
C	0.00000	0.50000	0.50000	0.70711			
E	0.00000	0.00000	0.00000	0.00000	1.00000		
E	0.00000	0.25000	0.25000	0.35355	0.50000	0.70711	
F	0.00000	0.50000	0.00000	0.00000	0.50000	0.00000	0.70711

Fig. A-3. Lower triangular L matrix stored in the output SAS® data set SAVE.L for the example problem.

Appendix B-1

Appendix B-1. SAS/IML™ program MAKEL.SAS.

```
1       * PROGRAM TO CREATE L;
        * WRITTEN BY R.A. MCLEAN AND D.M. PANTER;
        * TAKEN AFTER HENDERSON'S 1975 PAPER DESCRIBING THE ALGORITHM;

5       LIBNAME SAVE 'C:\BLUP';
        FILENAME INDEX 'C:\BLUP\INDEX.SAS';
        OPTION LS = 72;

        * CALL INDEX.SAS PROGRAM;
10      %INC INDEX;
        * RETURNS DATA SET NAMED A WHICH IS COMPLETELY NUMERIC;

        * INITIALIZE SAVE.L AND SAVE.PEDIGREE;
        DATA SAVE.L;
15              X=1;
        RUN;

        *       SAVE.PEDIGREE IS A DATA SET WHICH STORES THE NUMERIC
                PEDIGREE LIST.  CAN BE USED LATER IF DATA ARE APPENDED TO L.
20      ;

        DATA SAVE.PEDIGREE;
                FEMALE=0; MALE=0; OFF=1;
        RUN;
25
        PROC IML; RESET FW= 7;
                P=1;

                * READ IN VARIABLES FROM A (SUBSET OF SAVE.PEDLIST);
30              USE A;
                        READ ALL VAR {INDEX} INTO NNN;
                USE A;
                        READ ALL VAR {FEMALE MALE OFF} INTO PARE;

35
                * SAS/IML RENDITION OF HENDERSON'S 1975 ALGORITHM TO MAKE L;
                START AA;
                        DO II = 1 TO NROW(PARE);
                                DON=J(2,1,0);
40                              PAR =PARE[II,1:2];
                                NN=NNN[II,1];
                                LAPP=J(NN,1,0);
                                UPDATE=PARE[II,];
                                EDIT SAVE.PEDIGREE;
45                                      READ ALL VAR 'OFF' INTO PARENT;

                                DO I = 1 TO NROW(PARENT);
                                        IF PAR[1,1] = PARENT[I,1]  THEN DO;
                                                        DON[1,1] = I;
```

```
50                      END;

                        IF PAR[1,2] =  PARENT[I,1]  THEN DO;
                                       DON[2,1] = I;
                        END;
55                  END;

                    P=MIN(DON);Q=MAX(DON);
                    IF Q = 0 THEN GOTO ZERO; * NO PARENTS;
                    IF P = Q THEN GOTO SAME; * SAME PARENTS;
60                  IF P < Q THEN IF P > 0 THEN GOTO TWO; * TWO
                    PARENTS;
                    IF P = 0 THEN  DO;
                           P = Q;
                    END;

65                  GOTO ONE;

                    * SUBROUTINE IF ONE PARENT IS KNOWN;
                    ONE:
                           PPOINT1=1+P*(P-1)/2;
70                         PPOINT2=P*(P+1)/2;

                           PPPP=(DO(PPOINT1,PPOINT2,1));

                           EDIT SAVE.L;
75                                     READ INTO X POINT PPPP;
                           CLOSE SAVE.L;

                           SUM2=0;

80                         DO J=1 TO P;
                                       LAPP[J,1]=X[J,1]/2;
                                       SUM2=SUM2 + LAPP[J,1]**2;
                           END;

85                         LAPP[NN,1] = SQRT(1 +  - SUM2);

                           * UPDATE MATRICES;
                           EDIT SAVE.L;
                                       APPEND FROM LAPP;
90                         CLOSE SAVE.L;

                           EDIT SAVE.PEDIGREE;
                                       APPEND FROM UPDATE;
                           CLOSE SAVE.PEDIGREE;
95
                           GOTO LAST;

                    * SUBROUTINE IF TWO PARENTS ARE KNOWN;
                    TWO:
```

```
100                 PPOINT1=1+P*(P-1)/2;
                    PPOINT2=P*(P+1)/2;
                    QPOINT1=1+Q*(Q-1)/2;
                    QPOINT2=Q*(Q+1)/2;
                    QQQ=(DO(PPOINT1,PPOINT2,1)) ||
105                 (DO(QPOINT1,QPOINT2,1));

                    EDIT SAVE.L;
                                READ INTO X POINT QQQ;
                    CLOSE SAVE.L;
110
                    SUM1=0;
                    SUM2=0;
                    PP1=P+1;

115                 DO J=1 TO P;
                                LAPP[J,1]=(X[J,1]+X[(P+J),1])/2;
                                SUM1=SUM1+X[J,1]*X[(P+J),1];
                                SUM2=SUM2 + LAPP[J,1]**2;
                    END;
120
                    DO J=(PP1) TO Q;
                                LAPP[J,1]=X[(P+J),1]/2;
                                SUM2=SUM2 + LAPP[J,1]**2;
                    END;
125
                    LAPP[NN,1] = SQRT(1 + .5*SUM1 - SUM2);

                    * UPDATE MATRICES;
                    EDIT SAVE.L;
                                APPEND FROM LAPP;
130                 CLOSE SAVE.L;

                    EDIT SAVE.PEDIGREE;
                                APPEND FROM UPDATE;
                    CLOSE SAVE.PEDIGREE;
135
                    GOTO LAST;

                * SUBROUTINE IF BOTH PARENTS ARE THE SAME
                (INBRED);
                * NOT INCLUDED IN HENDERSON'S FORTRAN PROGRAM;
140             SAME:
                    PPOINT1=1+P*(P-1)/2;
                    PPOINT2=P*(P+1)/2;
                    QQQ=(DO(PPOINT1,PPOINT2,1));
145
                    EDIT SAVE.L;
                                READ INTO X POINT QQQ;
                    CLOSE SAVE.L;
```

```
150                         SUM1=0;
                            SUM2=0;
                            PP1=P+1;

                            DO J=1 TO P;
155                                 LAPP[J,1]=(X[J,1]);
                                    SUM1=SUM1+X[J,1]**2;
                                    SUM2=SUM2 + LAPP[J,1]**2;
                            END;

160                         LAPP[NN,1] = SQRT(1 + .5*SUM1 - SUM2);

                            * UPDATE MATRICES;
                            EDIT SAVE.L;
                                    APPEND FROM LAPP;
165                         CLOSE SAVE.L;

                            EDIT SAVE.PEDIGREE;
                                    APPEND FROM UPDATE;
                            CLOSE SAVE.PEDIGREE;
170
                            GOTO LAST;

                        * SUBROUTINE IF NO PARENTS ARE KNOWN;
                        ZERO:
175                             LAPP[NN,1] = 1;

                                EDIT SAVE.L;
                                        APPEND FROM LAPP;
                                CLOSE SAVE.L;
180
                                * UPDATE MATRICES;
                                EDIT SAVE.PEDIGREE;
                                        APPEND FROM UPDATE;
                                CLOSE SAVE.PEDIGREE;
185
                                GOTO LAST;

                        STOP:
                                INFO='TOO MANY PARENTS';
190                             PRINT INFO;
                                GOTO LAST;

                        LAST:
                                JUNK = 1;
200
                END;

        FINISH;

205     RUN AA;
```

```
            QUIT;
209     /* END PROGRAM THAT CREATES AND STORES THE L MATRIX */
```

Appendix B-2

Appendix B-2. SAS/IML™ program INDEX.SAS.

```
1       *       SAS/IML PROGRAM TO CONVERT CHARACTER PEDIGREES INTO
                NUMERIC PEDIGREES.  USES THE VALUE OF INDEX IN SAVE.PEDLIST
                AND THE IDENTIFICATION NUMBER FOR EACH INDIVIDUAL.  WRITTEN
                BY D.M. PANTER.
5       ;

        OPTIONS LS=72;
        LIBNAME SAVE 'C:\BLUP';

10      * READ IN SAVE.PEDILIST;
        DATA PED; SET SAVE.PEDLIST1;

        PROC IML; RESET NOPRINT;
                USE PED;
15                      READ ALL VAR {FEMALE MALE OFF} INTO PED
                                (|COLNAME=VARNAME|);
                USE PED;
                        READ ALL INTO INDEX;
                CLOSE PED;
20
                SIZE=NROW(PED);

                FEMALE=J(SIZE,1,0);
                MALE=J(SIZE,1,0);
25
                * MODULE TO MATCH INDEX NUMBERS THROUGH THE WHOLE DATA
                SET;
                START MATCH;

                DO I=1 TO SIZE;
30                      DO J=I TO SIZE;
                                IF PED(|J,1|)=PED(|I,3|) THEN
                                        FEMALE(|J,|)=INDEX(|I,|);
                                IF PED(|J,2|)=PED(|I,3|) THEN MALE(|J,|)=INDEX(|I,|);
                        END;
                END;
35              NEWPED=FEMALE || MALE || INDEX || INDEX ;
                FINISH;

                RUN MATCH;

40              PRINT NEWPED;
                VARNAME=NAME(FEMALE,MALE,OFF,INDEX);

        * CREATE DATA SET A TO SEND BACK TO MAKEL.SAS;
        CREATE A FROM NEWPED (|COLNAME=VARNAME|);
```

```
45      APPEND FROM NEWPED;
        CLOSE A;
        QUIT;
49      RUN;
```

Appendix B-3

```
1       *       SAS/IML PROGRAM TO CREATE BIG A MATRIX.  THE PURPOSE OF THIS
                PROGRAM IS TO LIMIT THE INDIVIDUALS IN THE A MATRIX TO ONLY
                THOSE WHICH MIGHT BE USED IN LATER ANALYSES.  WRITTEN BY
                D.M. PANTER.
5       ;

        LIBNAME SAVE 'C:\PAPER4';
        OPTIONS LS=72;
        DATA A; SET SAVE.PARENTS;
10
        PROC IML;
                RESET FW=4;

                * READ IN LIST OF POTENTIAL INDIVIDUALS;
15              USE A;
                        READ ALL VAR {PARENT} INTO PARENT;
                CLOSE A;

                * FIND ALL POTENTIAL INDIVIDUALS IN SAVE.PEDLIST;
20              USE SAVE.PEDLIST VAR {OFF INDEX};
                        READ ALL VAR {INDEX} WHERE(OFF=PARENT) INTO B;
                                PRINT B;
                        C=B-1;
                        READ POINT C VAR {OFF} INTO NAME;
25              CLOSE SAVE.PEDLIST;


        MATTRIB B
                ROWNAME=NAME;
30      ;
        PRINT B;

        CREATE INDEX FROM B;
                APPEND FROM B;
35      CLOSE INDEX;

        SIZE=NROW(B);
        MAX=MAX(B);

40      * SET UP DIMENSIONS OF THE SMALL L MATRIX;
        LSMALL=J(SIZE,MAX,0);

        *       MODULE TO LOCATE IN SAVE.L THE ELEMENTS CORRESPONDING TO
                THE LIST OF POTENTIAL INDIVIDUALS FROM SAVE.PARENTS;
45
        START LOCATE;

                USE SAVE.L;
```

```
50          DO I=1 TO SIZE;
                START = (B(|I,|)-1) * B(|I,|) / 2 +1;
                STOPP = (START + B(|I,|) - 1);
                P=START:STOPP;
                READ POINT P INTO ROW;
55              DO J=1 TO B(|I,|);
                    LSMALL(|I,J|) = ROW(|J|);
                END;
            END;

60          CLOSE SAVE.L;

    FINISH;

    RUN LOCATE;
65
    *       L TIMES L TRANSPOSE EQUALS A MATRIX;
    A = LSMALL * LSMALL`;

    MATTRIB A
70          ROWNAME=NAME
            COLNAME=NAME
    ;


75  CREATE SAVE.BIGA FROM A;
            APPEND FROM A;
    CLOSE SAVE.BIGA;

    CREATE NAMES FROM NAME;
80          APPEND FROM NAME;
    CLOSE NAMES;
    DATA INDEX; SET INDEX; RENAME COL1=INDEX;
    DATA NAMES; SET NAMES; RENAME COL1=ENTRY;

85  * WRITE ENTRY, INDEX, AND A MATRIX TO SAVE.BIGA;
    DATA SAVE.BIGA; MERGE NAMES INDEX SAVE.BIGA; PROC PRINT; RUN;

88  /*  END PROGRAM THAT MAKES BIG A MATRIX  */
```

Appendix B-4

Appendix B-4. SAS/IML™ program PREDICT.SAS.

```
1       *       SAS/IML PROGRAM TO MAKE BLUP FROM YIELD DATA IN SAVE.YIELD
                AND GENETIC RELATIONSHIPS STORED IN SAVE.BIGA.  WRITTEN BY
                D.M. PANTER.
        ;
5       LIBNAME SAVE 'C:\BLUP';
        LIBNAME TEMP 'C:\BLUP';

        * READ IN SAVE.YIELD;
        DATA A;
10              SET SAVE.YIELD;
                KEEP ENTRY ENV YIELD;
                * NOTE: THIS DATA SET MUST HAVE ENTRY, ENV, AND YIELD;
        RUN;
        PROC IML;
15      RESET FW=4;

        * READ IN THE BIG A MATRIX;
        USE SAVE.BIGA;
                READ ALL VAR {ENTRY} INTO ENTRY;
20              READ ALL INTO INDEXG;
        CLOSE SAVE.BIGA;


        SIZE=NCOL(INDEXG);
25      INDEX=INDEXG[,1];
        BIGG=INDEXG[,2:SIZE];

        * READ IN LIST OF PARENTS IN YIELD DATA SET;
        USE A VAR {ENTRY};
30              READ ALL VAR {ENTRY} INTO PARENT;
        CLOSE A;

        * LOCATE PARENTS IN BIG A CORRESPONDING TO YIELD DATA SET;
        USE SAVE.BIGA;
35              FIND ALL WHERE (ENTRY=PARENT) INTO POINTER;
                READ ALL VAR {ENTRY} WHERE (ENTRY=PARENT) INTO NAME;
                READ ALL VAR {INDEX} WHERE (ENTRY=PARENT) INTO SINDEX;
        CLOSE SAVE.BIGA;

40

        /*
        MODULE WHICH BUILDS SMALL A MATRIX FROM BIG A MATRIX.  ONLY
        ENTRIES IN SAVE.YIELD ARE EXTRACTED FROM THE BIG A MATRIX.
        */
45
        START BUILDG;
        SGSIZE=NROW(POINTER);
        SMALLG=J(SGSIZE,SGSIZE,0);
```

```
50      DO I=1 TO SGSIZE;
            DO J=1 TO SGSIZE;
                SMALLG[I,J]=BIGG[(POINTER[I]),(POINTER[J])];
            END;
        END;
55
        FINISH;
        RUN BUILDG;

        * CREATE A DATA SET WITH ENTRY AND INDEX;
60      CREATE ELIST VAR {NAME SINDEX};
            APPEND;
        CLOSE ELIST;
        MATTRIB SMALLG
            ROWNAME=NAME
65          COLNAME=NAME
        ;


        * OUTPUT SMALL A TO DATA SET;
70      CREATE TEMP.SMALLA FROM SMALLG;
            APPEND FROM SMALLG;
        CLOSE TEMP.SMALLA;

        * MERGE INDEX LIST WITH DATA AND SORT;
75      DATA ELIST;
            SET ELIST;
            RENAME NAME=ENTRY;

        PROC SORT DATA=ELIST;
80          BY ENTRY;
        PROC SORT DATA=A;
            BY ENTRY;
        DATA A;
            MERGE A ELIST;
85          BY ENTRY;
            IF SINDEX=. THEN DELETE;
        PROC SORT;
            BY SINDEX ENV;

90
        * DELETE MISSING OBSERVATIONS BEFORE READING XNAMES;
        * DATASET B IS USED ONLY TO READ IN XNAMES;
        DATA B; SET A;
            IF YIELD=. THEN DELETE;
95

        * FORM DESIGN MATRICES;
        PROC IML;
            RESET FW=4;
```

```
100
      * READ THE G MATRIX, INVERT, AND MULT. BY H CONSTANT;
      USE TEMP.SMALLA;
            READ ALL INTO G;
      CLOSE TEMP.SMALLA;
105
      * MULTIPLY A INVERSE BY H2/(1-H2);

      GINV=INV(G) * 1.5;  *NOTE HERITABILITY OF 0.4;

110   * READ PERFORMANCE DATA;
      USE A;
            READ ALL VAR {SINDEX} INTO RANDOM;
            READ ALL VAR {ENV} INTO FIXED;
            READ ALL VAR {YIELD} INTO Y;
115   CLOSE A;

      *PRINT RANDOM FIXED Y;

      Z=DESIGN(RANDOM); OLDZ=Z;
120
      * LOOK FOR MISSING VALUES AND ADJUST MATRICES ACCORDINGLY;

      * IF PERFORMANCE DATA IS MISSING, CHANGE Z TO A COLUMN OF ZEROS;
      DO I=1 TO NROW(Y);
125         IF Y[I]=. THEN DO;
                  Z[I,]=0;
            END;
      END;

130
      * LOCATE NON-MISSING RECORDS AND FORM NEW MATRICES;
      THERE=LOC(Y);

      NEWENV=J(NCOL(THERE),1,0);
135   NEWY=J(NCOL(THERE),1,0);
      NEWZ=J(NCOL(THERE),NCOL(Z),0);
      DO I=1 TO NCOL(THERE);
            NEWENV[I,]=FIXED[(THERE[I])];
            NEWY[I,]=Y[(THERE[I])];
140         NEWZ[I,]=Z[(THERE[I]),];
      END;


      MU=J(NROW(NEWY),1,1);
145   X=MU || DESIGN(NEWENV);
      Y=NEWY;
      Z=NEWZ;


150   * GET NAMES OF RANDOM AND FIXED EFFECTS;
```

```
      NO_FIXED=NCOL(X)-1;  XLOC=J(NO_FIXED,1,0);
      NO_RAN=NCOL(Z);  ZLOC=J(NO_RAN,1,0);


155   DO I=1 TO NO_RAN;
            TEMP=OLDZ[,I];
            AA=LOC(TEMP);
            ZLOC[I,]=AA[1,1];
      END;
160
      * READ IN NAMES OF ENTRIES;
      USE A;
            READ POINT ZLOC VAR{ENTRY} INTO ZNAMES;
      CLOSE A;
165
      DO I=1 TO NO_FIXED;
            TEMP=X[,I+1];
            AA=LOC(TEMP);
            XLOC[I,]=AA[1,1];
170   END;

      * READ IN NAMES OF ENVIRONMENTS. NOTE THAT B HAS NO MISSING OBS;
      USE B;
            READ POINT XLOC VAR{ENV} INTO XNAMES;
175   CLOSE B;

      MEAN="MEAN";
      SOLLAB=MEAN // CHAR(XNAMES)  // ZNAMES;

180   * CLEAN UP UNNEEDED MATRICES;
      FREE RANDOM FIXED XLOC ZLOC;


      * FORM MATRICES OF THE BLUP EQUATIONS;
185   UL=X'*X;
      UR=X'*Z;
      LL=UR';
      LR=(Z'*Z) + GINV;

190   LHS=(UL || UR) // (LL || LR);
      RHS=(X'*Y) // (Z'*Y);

      * CLEAN UP UNNEEDED MATRICES;
      FREE UL UR LL LR;
195
      * SOLVE FOR THE SOLUTION VECTOR;
      CC=SWEEP(LHS);
      SOL=CC * RHS;

200   MATTRIB SOL ROWNAME=SOLLAB;
```

```
      PRINT SOL;

      * SET UP K MATRICES TO MAKE ESTIMATES OF MEAN AND SE;
205   FIXEDL=J(1,NO_FIXED,(1/NO_FIXED));
      MU=1;
      FIXEDL=REPEAT((MU || FIXEDL),NO_RAN,1);
      RANL=I(NO_RAN);

210   K=FIXEDL || RANL;


      EST=K * SOL;
      TEMP=VECDIAG(K * CC * K');
215   STDERR=SQRT(TEMP);
      FINAL=EST || STDERR;
      MATTRIB FINAL
            ROWNAME=ZNAMES;

220
      * WRITE PREDICTIONS TO DATA SET;
      CREATE PREDICT VAR {ZNAMES EST STDERR};
            APPEND;
      CLOSE PREDICT;
225
      * RENAME VARIABLES;
      DATA PREDICT;
            SET PREDICT;
            RENAME ZNAMES=ENTRY EST=BLUP;
230
      * SORT BY DESCENDING YIELD;
      PROC SORT DATA=PREDICT NODUPKEY OUT=FINAL;
            BY ENTRY;
      PROC SORT;
235         BY DESCENDING BLUP;
      PROC PRINT;
      RUN;

239   /*      END BLUP PREDICTION PROGRAM  */
```

PART VI

Overall Summary

CHAPTER 1

Summary of Parts II-V

Progressive plant breeders are evaluating constantly new approaches to estimating genotypic means that may help increase their success in developing new cultivars. Using mixed linear models to make best linear unbiased predictions (BLUP) of genotypes had been proven effective for predicting breeding values of dairy sires based on the milk yield performances of their dams and daughters. The purpose of this research project was to determine if the same methodology utilized in dairy cattle sire evaluation also could be employed effectively for predicting seed yield of soybeans. Based on the results described in the preceding papers, this methodology can have very broad and effective application in plant breeding.

The first paper described how BLUP compared to a classical method of estimation, midparent value (MPV), for evaluating seed yield means of future crosses based on historical performance data of the potential parents. When adequate data were available for all potential parents, BLUP provided more precise predictions of future crosses than estimates from MPV. The superiority of BLUP became much more apparent when data were limiting or unavailable for some of the

potential parents. Under these conditions, BLUP provided predictions of future crosses more precisely than MPV did under the best conditions (*i.e.* adequate data available on all potential parents). The main reason for the superiority of BLUP is in its use of known genetic relationships among individuals for which data are available. BLUP uses these relationships to increase the effective number of observations on individuals for which little or no data are available. Since limited parental data is generally the rule for plant breeders, BLUP methodology should greatly increase the efficiency of choosing parents with the highest probability of producing superior progeny by providing better estimates of the breeding values of potential parents.

Since BLUP utilizes information on relatives of an individual to predict the individual itself, breeders need to know how much data for relatives are needed to predict precisely the value of the individual. The second paper described how progeny and grand-progeny of an individual affected its breeding value when data for the individual *per se* were unavailable, limited, and abundant. Results from these analyses showed that when data were unavailable for a potential parent, information from progeny of the parent, in sufficient quantity, provided precise estimates of the breeding value of a parent using BLUP. Likewise, when limited data were available for a parent and ample data were available from its progeny, predicted breeding value were more precise than when there were no data on the parent *per se*. The same results were observed when grand-progeny data were used to predict the breeding value of an individual. As expected, data from grand-

progeny were less effective than data from progeny for predicting breeding values of parents, but grand-progeny data could provide fairly precise predictions if these data were the only data available to the breeder.

After crosses are made, new genotypes from those crosses must be evaluated so that the breeder can identify the superior genotypes from among a larger group. In the early stages of performance testing, many individuals which are highly related to one another are evaluated together in order to distinguish which genotypes within the group should be selected for further evaluation. Seed supply and/or cost constraints usually dictate the number of trials the breeder can use to evaluate new material. Since it had been demonstrated previously that BLUP could be used to provide precise predictions of an individual based on performance of its relatives by increasing the effective number of observations on the individual, BLUP should also be effective in predicting means of new genotypes in the early stages of evaluation. In the third paper, BLUP was compared to the standard method of estimation, BLUE, for evaluating the mean yield of 24 late generation soybean crosses based on yield trials in one and/or two environments. In almost every case, predictions of the yield performance of the 24 crosses were more precise using BLUP. It was demonstrated also that inclusion of historical records on the direct parents of the 24 crosses did little to enhance the precision of predictions using BLUP. However, some of the 24 crosses used in this study were highly related and affected the predictions of one another to a large extent, nullifying the value of the parental data. Including historical parental data might have been more valuable if the crosses being

evaluated had been more genetically diverse. These results indicated that BLUP is superior to BLUE for predicting mean seed yield of soybean crosses in the early stages of performance testing. Because the genetic material used in this study might have created a bias against using historical parental data, it was recommended that parental data, if available, should be utilized through BLUP to enhance predictions of their progeny.

Based on the results described above, it was apparent that BLUP could be utilized for many applications in a plant breeding program. However, employing the methodology would be difficult for many plant breeders because of the complex computing rituals involved in making predictions. The fourth and final paper described the computing strategy used to make predictions from BLUP. Using an example model, details of the input to and output from computer programs written to perform BLUP analysis were described. Source code for all computer programs was also provided. By furnishing the details of these analyses, the necessary base of information in both theory and application was provided which might encourage plant breeders to utilize this potentially valuable methodology.

Vita

The author, Donald McLean Panter, Jr., was born in Atlanta, Georgia, on June 5, 1962. He attended elementary schools in Tennessee, North Carolina, and Oklahoma. He graduated from Lejeune High School, Camp Lejeune, North Carolina, in 1980. Upon graduation, he enrolled at the University of Tennessee, Knoxville, in September of 1980. In December 1984, he received a Bachelor of Science Degree in Agriculture with Honors, with a major in Plant and Soil Science.

In January 1985, he entered the Graduate School of the University of Tennessee and worked in the area of Plant Breeding. In October 1985, he accepted a position with the Department of Plant and Soil Science as a Senior Research Assistant in Soybean Breeding. In August 1987, he was awarded a Master of Science Degree in Agriculture with a major in Plant and Soil Science.

In September 1987, he began working toward a doctoral degree in Plant and Soil Science with a major in plant breeding at the University of Tennessee, Knoxville. In July 1989, he was promoted to Research Associate and continued to work in the area of soybean breeding. In August 1991, he was awarded a Doctor of Philosophy in Agriculture, with a major in Plant and Soil Science and a minor in Statistics.

Dr. Panter is the son of retired Marine Corps Master Gunnery Sergeant and Mrs. Donald M. Panter, Sr. of Turtletown, Tennessee. He is married to Jennifer

Lee Sanders Panter and has two children, Bonny Kaitlin and Donald McLean III.