Louisiana State University
LSU Digital Commons

LSU Historical Dissertations and Theses | Graduate School

2000

Biometrical Models for Predicting Future Performance in Plant Breeding.

Monica Graciela Balzarini
Louisiana State University and Agricultural & Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_disstheses

Recommended Citation
Balzarini, Monica Graciela, "Biometrical Models for Predicting Future Performance in Plant Breeding." (2000). *LSU Historical Dissertations and Theses*. 7178.
https://digitalcommons.lsu.edu/gradschool_disstheses/7178

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Historical Dissertations and Theses by an authorized administrator of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

BIOMETRICAL MODELS FOR PREDICTING FUTURE PERFORMANCE IN PLANT BREEDING

A Dissertation

Submitted to The Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Agronomy

By
Mónica Graciela Balzarini
B.S., Universidad Nacional de Córdoba, Argentina, 1984
M.Sci. in Biometry, Universidad de Buenos Aires, Argentina, 1995
May, 2000

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: 9979242

UMI

UMI Microform 9979242
Copyright 2000 by Bell & Howell Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

To Walter, Carla, José and Federico

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ACKNOWLEDGEMENTS

I am specially grateful to Dr. Scott Milligan whose own work, based on continuous plant breeding and genetic research, has clarified my thinking. He provided excellent guidance and support for the realization of this study. I deeply appreciate the constructive advice from Dr. Lynn LaMotte, Dr. Luis Escobar, Dr. Manjit Kang, and Dr. Brad Venuto, members of my advisor committee. My sincere thanks to the faculty, staff, and fellow graduate students I met at Louisiana State University for creating a friendly and constructive atmosphere so encouraging during these years. I must recognize my friends and work mates from Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Argentina, for their continuous support throughout the years I postponed my activities in that Institution. My gratitude to Universidad Nacional de Córdoba and FOMEC-Ministerio de Educación de la Nación Argentina for their financial support.

I wish to thank Walter, my husband, for offer me a great companion in all project I have started, and I must especially thank my three kids for their patience and enthusiastic support. I hope this work benefit them.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE OF CONTENTS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

The plant breeding process begins with the selection of parents and crosses. Promising progeny from these crosses progress through a series of selection stages that typically culminate in multi-environment trials. I evaluated best linear unbiased predictors (BLUP), other predictors and prediction models at the initial (cross prediction), early replicated testing and late (multi-location) stages of a sugarcane breeding selection cycle. Model and predictor accuracy was assessed in the first two stages by using cross-validation procedures. I compared statistical models of progeny test data in their ability to predict the cross performance of untested sugarcane crosses. Random parental effect predictors and a random cross effect predictors were compared to mid-parent values (MPV) derived from a fixed female-male parental effect model. The cross effect model was evaluated with and without incorporating the genetic relationships among tested crosses into the BLUP derivation. Models with BLUP-based predictors showed smaller mean square prediction error and higher fidelity of top cross identification than the MPV for all traits evaluated. The MP-BLUP was consistently the best one.

Prediction of *per se* (genotype) performance is needed during the selection process and requires combining information from different trials. The study investigated three mixed models involving three versions of BLUPs estimated under different strategies, a fixed least squares genotype means model, and four check-based methods for combining information at early replicated stages. BLUP-based predictors

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

were superior to the currently used predictor (average percent of check cultivar). In addition, BLUP accuracy was not dependent on check values.

In later selection stages, when few and highly selected genotypes are evaluated, genotype effects may be assumed fixed. By assuming genotype-by-environment interaction effects as random, the modeling of the covariance matrix allowed direct estimation of stability and genotype-by-environment measures. Closely related mixed models involving covariance parameters related with genotype-by-environment interaction were estimated. The covariance structure of the observations under the mixed models adjusted the genotype mean separation. Stability parameters were integrated into broad (across environment) and narrow (environment specific) inferences about genotype yield performances. A procedure to obtain visual representation of the genotype-by-environment interaction (BIPLOT) under a mixed AMMI model was also derived.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 1

PLANT BREEDING DATA AND MIXED MODELS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. INTRODUCTION

1.1. Plant Breeding and Statistical Modeling

The ultimate goal of plant breeding is to generate productive cultivars improved for one or more traits. The breeding process begins with the selection of parents that possess desired attributes. Parents are typically derived from advanced stages of selection or they are recognized commercial lines or cultivars. Hybridization of these parents generates progeny that are typically screened via a series of selection stages. The choice of parents and hybrid combinations affect the quality of the progeny. The selection process at the initial stages of the breeding process, when there is a large amount of new material, typically uses small unreplicated plots or only limited replication. As progressive selection for desired traits reduces the size of the progeny population, breeders collect more objective data and use larger plots and more replications. Researchers commonly use replicated multi-environment trials in the final stages of the selection process.

Phenotypic data are generated at each stage in which the genotypes are tested. The data can be analyzed for purposes such as parent selection, ranking of genotypes for progressive advancement in stages, and comparing performance of advanced genotypes in different environments. The data can also be used to diagnose the population and prescribe the most appropriate strategies to maximize progress toward short and long-term breeding goals.

When analyzing cross performance to select appropriate parental combinations for crossing, it is important to note that parental generations are rarely discrete. They

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

commonly originate from different selection and crossing series. Because the genotypes that represent potential parents often derive from different stages of selection, the amount and precision of the data may dramatically vary. *Per se* variety (parent) evaluation usually requires several years (Brown and Dale, 1998). Selecting superior genotypes in the early generations might be highly ineffective for several crops, especially those clonally propagated (Skinner, 1971; Caligari et al., 1986; Gopal et al., 1992). Several researchers have demonstrated the gain in efficiency of selection by using cross prediction trials or progeny tests for family selection (Hogarth, 1971; Brown et al., 1988; Zaunbrecher 1995; Cox et al., 1996; Simmonds, 1996). Thus, cross appraisal or progeny tests are commonly employed at the beginning of each breeding-cycle (Milligan and Legendre, 1991; Cox and Hogarth, 1993). However, only a few of all potential hybrid combinations are actually made and evaluated in progeny trials.

Typically, the performance of a new or untested cross combination is predicted by calculating mid-parent values (MPV) of the raw or scaled parental mean. These means are based on observed (*per se*) records of potential parents. Improved estimates of parental means for a trait are often obtained with some form of an additive linear model. Such models adjust observed values for non-genetic effects to obtain better estimates of the genetic effects (Panter and Allen, 1995). A classic method of obtaining parental genetic effects is by combining data across progeny tests and considering all effects in the model as fixed (White et al., 1986). Unfortunately, cross appraisal databases are typically incomplete and unbalanced, which creates theoretical concerns about the fixed linear model underlying the mid-parent value prediction (Henderson, 1973).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Mixed models provide alternative analytical approaches that may overcome limitations of the fixed analytical approach (Henderson, 1974, 1975). Best Linear Unbiased Prediction (BLUP), as it is presented in a mixed linear model framework (Henderson, 1975; Searle et al., 1992), has been used for prediction and estimation of genetic merit of tested material in plant breeding (Bridgess, 1989; Chang and Milligan, 1992; Chang, 1996; Cox and Stringer, 1998). Mixed model-based prediction has also been proposed for predicting the performance of untested crosses in the production of hybrid crops such as corn (*Zea mays* L.) (Bernardo, 1994) and soybean (*Glycine max* L. Merr.) (Panter and Allen, 1995). This method demonstrates better prediction accuracy than that obtained by using a fixed linear model. The mixed model prediction of untested crosses relies on the genetic relationship between tested and untested crosses.

Efficient selection and advancement of individual genotypes from one stage to the next assumes the current data predicts the future *per se* (genotype) performance. It ideally uses all the information that is available from past trials. Yet the trials from different stages vary in number of entries, plot size, replication, genetic level of elitism, and experimental precision. To overcome these incongruities among trials and stages, breeders often incorporate commercial check cultivars into the tests. Typically, experimental entry values are expressed relative to the check(s) (Hill and Rosenberg, 1985). But, usually with time, the checks are changed as the experimental population exceeds the performance of older checks or when new and more relevant commercial checks become available. Commercial checks are individual genotypes or cultivars that generally vary in performance among trials and themselves. Thus, the use of checks to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

combine information has limitations. Moreover, during the selection stages many related individuals are tested, yet classic analysis seldom incorporates the correlated information into individual performance predictions. Mixed linear model approaches may circumvent the problems of fixed ANOVA methods for combining information from different trials and incorporating genetic correlations by treating genotype effects as random variables (Stroup, 1989; Littell et al., 1996; Federer and Wolfinger, 1998).

Most agricultural and economically important traits of commercial crops are quantitative in nature, are controlled by polygenes with various kinds of genetic effects and are affected by the environment. Thus, the variety trials commonly conducted in the latest stages of a breeding cycle involve a few highly selected genotypes tested in several environments. Broad (across environments) inference, narrow (environment-specific) inference and genotype-by-environment interaction implications are important considerations (Milliken and Johnson, 1994; Littell et al., 1996; Kang and Gauch, 1996; Shafii and Price, 1998). The information related to variety trials is often incomplete over time since not all genotypes are evaluated in all environments. The genotype effects are seldom treated as random effects, whereas environments and/or genotype by environment interaction may be regarded as random. A random approach for environment and genotype-by-environment interaction effects allows the modeling of correlation structures throughout their associated variance components (Cullis et al., 1996; Magari and Kang, 1997; Piepho 1994, 1997, 1998a).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The Louisiana Agriculturai Experiment Station (LAES) sugarcane breeding program provides a good example of a common plant breeding data structure (Milligan, 1994).

Sugarcane (*Saccharum* spp.) is a clonally propagated crop where crosses among clones, used as female and male parents, are used to obtain new genetic material. The breeding program uses several sequentially planted selection stages to identify and select the best clones within each crossing series. Material with commercial potential is ultimately evaluated in yield trials in several environments (year and location combinations). Sugarcane is planted in Louisiana in late summer or early fall and is typically harvested three times (plant cane, first ratoon and second ratoon crops), once each fall, prior to fallowing for replanting. The program uses 10 selection stages, and requires 12 years from crossing to varietal release. The process typically begins by planting about 50,000 seedling progeny each year, representing 150 to 250 crosses among 70 to 80 parents (Table 1.1).

A replicated progeny test of crosses is planted each year. Family selection among progeny is based on objective data from the progeny test and is used to initially select the most promising families (crosses). Subsequently, individual plant selection is performed within the selected families. Stool weight, freedom from diseases, and hand-refractometer-Brix serve as the visual selection criteria for the single-plant selection stage. A subjective cane yield rating followed by selection for high Brix is used in the first clonal trials. Objective yield data through the second ratoon crop are collected from the second clonal stage and successive stages.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Replicated yield trials are initiated in the third clonal trial (Increase) and experimental clones might be first used as parents in year six. Replicated testing of no more than 10 experimental clones culminates in the advanced variety trials (Outfield). In any given year, populations exist in all stages of the selection program.

Table 1.1. Number of locations, replications and crops used in a typical year of the Louisiana Sugarcane Variety Development Program †

Year	Stage	Locations	Replications	Crops harvested	Total harvested yield plots available when planting next stage ‡
		number			
1	Crossing	-	-	-	-
2	Seedling	1	1	1	0
3	1st Clonal	1	1	2	0
4	2nd Clonal	1	1	3	0
5	Increase	2	1	3	0
6	Nursery	3	2	3	2
7	Infield/Nurs.	1/3	2/2	3/3	4
8	Infield	1	1	1	12
9	Outfield	10	3	3	28
10	Outfield	10	3	3	52
11	Outfield	10	3	3	90
12	Outfield	10	3	3	150
13	Release				240

† Abridged from Milligan, 1994, Table 1.

‡ First ratoon second clonal plots are harvested prior to planting the nursery plots. This is earlier than normal. Otherwise, plots are not harvested until after planting, hence there is a two-year delay in harvest information prior to planting.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The LAES database is characterized by a complex structure involving several types of genetic and experimental correlations. It contains incomplete and unbalanced information for each genotype, i.e., information on all genotypes is not available at the same selection stage and in the same trial. To make informed decisions, statistical models and estimation procedures that can effectively handle the database features are needed.

The particular characteristics of each stage in the breeding process demand different statistical modeling strategies. The general objective of this study was to compare biometrical models for three general stages of the breeding process. The first part compared models for predicting cross performance in the hybridization stage. The second part analyzed *per se* (genotype) prediction performance at early stages of the breeding process when considering genotype effect as a random variable might be convenient. The last part looked at predicting *per se* (genotype) performance at late stages of the breeding process when highly selected genotypes, usually assumed fixed, are evaluated across several environments.

The research explores mixed linear models to improve predictions of cross performance and genotype *per se* performance in a typical sugarcane breeding program. It analyzes models and performs estimation procedures of the underlying variance-covariance structures at three different stages of the Louisiana Sugarcane Variety Development program, i.e., crossing, selection stages, and advanced variety trials.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Identifying the best parents and cross combinations should improve the likelihood of producing elite progeny and selecting superior genotypes for potential release as cultivars. The prediction of *per se* performance would aid genotype selection across stages and enhance variety recommendations. By increasing the probability of selecting the best parents and lines, breeders may increase the selection intensity in early stages. Hence, better predictions in a breeding program may accelerate early stage selection and ultimately shorten and or enhance the effectiveness of the selection cycle.

1.2. Predicting Genotype Performance: The Modeling Strategy

Statistical modeling is based on the specification of the expected value and the variances and covariances of observed data. Predictions depend on that modeling. The conventional general linear model coupled with ordinary least squares estimation procedures (OLS), useful as it is in many experiments in agriculture, is too restrictive to perform satisfactory data analyses for the typical data structure of most breeding programs. Error structure in "real world" experiments is often more complex than used in standard linear models for conventional data analysis (Stroup, 1989).

In contrast, the general linear mixed model can accommodate covariance structure among observations. Standard linear models usually assume independence. The mixed model handles these correlations with random effects and their associated variance components, modeling variability over and above the component associated with residual error (Wolfinger and Tobias, 1998). Mixed linear model approaches can circumvent the troublesome ANOVA for handling unbalanced data and complex models.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Mixed model analysis applies particularly to research involving factors with a few levels that usually can be controlled by the researcher (fixed) as well as factors with levels that are beyond the researcher's control (random). These random factors vary from experiment to experiment, and may be interpreted in the context of a symmetric probability function. Most breeding trials have some mixed model aspect. The two parents of each hybrid variety contribute randomly by one half of its genetic make-up. The allelic complement passed on to its progeny is different for each descendent. The number of potential genotypes involved in the crossing process is large, but the number of realized effects is substantially less. Additionally, the distribution of genetic effects is reasonably symmetric for most important quantitative traits. Therefore, genetic effects may be reasonably assumed as random (Stroup, 1989; Henderson, 1990; Robinson, 1991). Federer (1997) commented on the random nature of genotypes in the early stages of a selection program.

However, at the later selection stages genotypes might be assumed as fixed since research is focused on a few selected genotypes. In the later selection stages, environmental and/or genotype by environment interaction effects may be considered random (Bridges, 1989; Piepho, 1994).

The mixed model framework is flexible enough to adjust to the structural changes and factors that affect the selection process during its different stages. It is generally applicable to a wide variety of quantitative genetics and breeding prediction problems. Mixed models have not been used in a unified framework in plant breeding.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Traditionally, mixed model applications in plant breeding have focused on population variance component estimation and identification of appropriate error terms to be used for testing fixed effect hypothesis (Cockerham, 1963; Falconer, 1989). Rarely have they have been used for the most general purpose of modeling the underlying covariance structure in the data.

Liang and Zeger (1986) and Zeger et al. (1988) discussed the interpretation of the mixed model estimates in both a subject-specific and population-average sense. A subject-specific approach focuses on the prediction of random effects for individuals and their relation to the population parameters (fixed parameters). With a population-average approach, the interest is primarily on fixed parameters. Variability arising from random effects is treated essentially as a nuisance parameter. The best linear unbiased predictor (BLUP), as a technique for predicting random effects (Harville, 1990; Robinson, 1991), should be understood as a subject-specific mixed model prediction (Wolfinger and O'Connell, 1993).

Henderson's work (1973,1974,1975) on BLUPs of genetic random effects in animal sciences represents the best known use of mixed model theory to predict future performance for breeding purposes. BLUPs of random genetic effects have been used for predicting genetic performance in crop plants on only a limited basis (White et al., 1989; Panter and Allen, 1995; Bernardo, 1994,1995,1999; Chang and Milligan, 1992; Piepho, 1994; Cox and Stringer, 1998).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The prediction value of unobserved or future performance is an important consideration in plant breeding. Prediction of random variable outcomes, in general, is a fundamental problem in statistics (Hinkley, 1979; Butler, 1986; Bjornstad, 1990). Assuming that there is *a priori* knowledge about the distribution of the parameters defining the variable to be predicted, predictions are obtained by finding the posterior distribution of the variable, given the data, from a Bayesian point of view (Gelman et al., 1995).

Besides the Bayesian approach to the prediction problem, the general mixed model allows prediction in a frequentist framework via the concept of conditional expectation without using a *priori* distribution. The conditional expectation of the random effects, given the observed data, is the BLUP of those random effects, and is also a Bayes estimator under normal priors (Robinson, 1991; Searle et al., 1992). Theoretically, BLUPs have the smallest mean squared error of prediction among all linear unbiased predictors, provided the assumed model holds and the parameters of the model are known (Searle et al., 1992). In practice, estimates replace parameters and different models for the variance-covariance structure of the observations lead to different BLUPs. Thus, the term BLUP is quite general and a precise identification of the underlying model is needed to avoid confusion.

Applications of a more general mixed model framework to combine information, estimate fixed and random parameters, and improve *per se* performance prediction produce smaller prediction errors when compared with the ordinary least squares approach for analyzing agricultural experiments (Wolfinger et al., 1997).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Using mixed model equations and simple correlation structures, Hill and Rosenberger (1985) showed the efficiency of BLUP for combining information for germplasm evaluation. Similar BLUPs, i.e., assuming random effects as stochastically independent, were reported effective by Piepho (1994) for modeling multi-environment variety trials. Oman (1991) and Gogel et al. (1995) have shown how to fit models of complex variance-covariance structure to genotype by environment data. Magari and Kang (1997) used mixed model estimation to consider the interaction of individual genotypes with environments for stability analysis. Interaction-term variance components were estimated for each genotype by using the mixed model equations. The variance components were used as stability measures. They are equivalent to Shukla's stability variances (Shukla, 1972), but it is important to note that they were estimated as parameters of a mixed model. Piepho (1998) put different well-known stability measures (Kang and Gauch, 1996) into a unifying mixed model perspective.

Mixed models have been successfully used for recovering inter-effect information from experiments that use designs such as augmented and lattice designs (Federer, 1997). Federer and Wolfinger (1998) have shown that the expected error mean square for differences (contrasts) of means is smaller when random effect information is recovered than when it is ignored. The variance components estimated in the mixed model framework are themselves informative in breeding. Heritabilities and response to selection are obtained from the variance components of the mixed model without regarding data unbalance.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Random variable predictions that involve estimation of fixed and random effects may be obtained using appropriate BLUPs and treatment (fixed effects) means. Under a general mixed linear model, the predictions account for involved variance and covariances.

Mixed model background has been developed over many years (Anderson and Bancroft, 1952; Henderson, 1953, 1974, 1975; Scheffé, 1956; Hayman, 1960; Searle, 1971, 1987; LaMotte, 1973, 1988; Rao, 1973, 1988; Harville, 1976, 1977, 1990; McLean et al., 1991; Searle et al., 1992; Khuri, 1998). The new mixed linear model approaches offer opportunities for plant breeders to better deal with complex databases. However, mixed linear models have rarely been applied in plant breeding before software such as PROC MIXED (SAS Inst., 1996) became available to overcome the computational demands of this approach (Littell et al., 1996; Piepho 1998; Wolfinger et al., 1997; Federer and Wolfinger, 1998).

1.3. The Mixed Linear Model: General Overview

The mixed model contains fixed effects that determine the mean of the data and random effects to model variance and covariance. Several authors attempted to give general definitions for fixed and random effects (Scheffé, 1956; Searle, 1971; Stroup, 1989; Robinson, 1991; Searle et al., 1992). Most of them approached the problem from a theoretical frequentist point-of-view and did not develop a clear definition. Analytically, clear reasons may exist for treating a factor as random or fixed. If the factor has a large number of levels, which are related to some probability function, it may be best to treat the factor as random.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The researcher should use a BLUP if the prediction of specific levels for the random effects, in a particular experiment, have importance.

The general form of a linear mixed model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is a n vector of observable random variables (data), $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices, β is a p vector of effects parameters having fixed values, and $\mathbf{u}$ (random effects) and $\mathbf{e}$ (error terms) are unobservable random m and n vectors, respectively. Usually $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_q]$ where each $\mathbf{Z}_i$ represents the model design matrix for the i*th* random factor and $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_q]$ where $\mathbf{u}_i$ is a m_i random vector. Note that $m = \Sigma m_i$. Assumptions about $E(\mathbf{u})$, $E(\mathbf{e})$, $\mathbf{G}$ -the variance-covariance matrix of the random effects in $\mathbf{u}$-, $\mathbf{R}$ -the variance-covariance matrix of the random error terms in $\mathbf{e}$-, and the covariance between $\mathbf{u}$ and $\mathbf{e}$ will define a particular mixed model.

When the vector of observations is normally distributed, the probability distribution of the data is completely determined by its mean and the variance-covariance matrix. The typical assumption of independence made in the general linear model is eliminated in the mixed model by modeling statistical correlations through $\mathbf{V}$, which is the matrix containing the variances and covariances of each observation. Models for the variance-covariance of the data, $\mathbf{V}$, are obtained by specifying the structure of $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{R}$.

Non-constant variance and covariance for both the random effects and the residual errors, as well as dependence of the variances on the levels of fixed and random factors can be introduced throughout different structures of the variance-covariance matrix $\mathbf{V}$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

A simple, yet important class of linear mixed models contains only one source of random effects and assumes $E(\mathbf{u}) = \mathbf{0}$, $E(\mathbf{e}) = \mathbf{0}$, $\mathbf{G} = \sigma_u^2 \mathbf{I}$, $\mathbf{R} = \sigma_e^2 \mathbf{I}$, and $Cov(\mathbf{u},\mathbf{e}) = \mathbf{0}$. The variance components, σ_u^2 and σ_e^2, are scalar-valued parameters. According to the previous assumptions, the expected value of the data is $E(\mathbf{y}) = \mathbf{X}\beta$, and the variance-covariance matrix is $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R} = \sigma_u^2 \mathbf{Z}\mathbf{Z'} + \sigma_e^2 \mathbf{I} = \sigma_e^2 (\mathbf{I} + \gamma \mathbf{Z}\mathbf{Z'})$, where $\gamma = \sigma_u^2 / \sigma_e^2$ is a variance component ratio. Note that the variance of the data is a linear function of the variance components.

In plant breeding, this model could be used when **y** contains the measured responses, such as plot yields from *m* genotypes, each represented by a random genetic effect $\mathbf{u} = [u_1,....,u_m]$ to be predicted, and β might be the vector of fixed parameters related to trial effects. The model assumption, $\mathbf{G} = \sigma_u^2 \mathbf{I}$, implies that the genetic effects are independent with zero mean and homogeneous variance denoted by σ_u^2. The model $\mathbf{R} = \sigma_e^2 \mathbf{I}$ for the variance-covariance of error terms implies that the error terms are uncorrelated with each other and that the residual variance is homogeneous, i.e. the same σ_e^2 for all observations. Note that the parametric function $\gamma / (1+\gamma) = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$ may be interpretable as a broad-sense heritability estimate for an individual plot-basis scenario (Nyquist, 1991). Thus, the simplest form for **G** and **R** is one that arises from independence and constant variances of the random effects and the error terms. However, the independence in the random effects does not imply that the observations are independent. The **V** matrix for this simple between-within mixed-model ANOVA is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

a block diagonal matrix indicating that observations within the same level of the random effects are equally correlated, and observations between different levels of the random effects are independent. Thus, if **Z** is a model matrix with 1s and 0s, each of the sub-matrices in the block diagonal of **V** will be a type of matrix with the property called compound symmetry (Jenrich and Schluchter, 1986). This is because all the diagonal elements are equal to $\sigma_u^2 + \sigma_e^2$, which is the variance of any observation, and the off-diagonal elements are equal to σ_u^2, which is the covariance between any pair of observations sharing the same random effect. Therefore, by considering **u** as random effects with variance $\mathbf{G} = \sigma_u^2 \mathbf{I}$, this sets up a common correlation among all observations having the same level of **u**.

The extension of the model to allow several random effects is straightforward. Assume that $q > 1$ random factors are considered in the model. For example, suppose family, plots within family, and rows within plots are all random sources of variation. If the levels within each factor are assumed to be independent and with homogeneous variance, say $\mathbf{G}_i = \sigma_{u_i}^2 \mathbf{I}$, and random effects are uncorrelated between sources then

$$E(\mathbf{y}) = \mathbf{X}\beta$$

$$\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R} = \sigma_{u1}^2 \mathbf{Z}_1\mathbf{Z}_1' + \ldots + \sigma_{u_q}^2 \mathbf{Z}_q\mathbf{Z}_q' + \sigma_e^2 \mathbf{I} = \sigma_e^2 (\mathbf{I} + \gamma_1 \mathbf{Z}_1\mathbf{Z}_1' + \ldots + \gamma_q \mathbf{Z}_q\mathbf{Z}_q')$$

where $\gamma_1, \ldots, \gamma_q$ represent the variance ratios for the respective random sources.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Different models may be assumed for the variance-covariance matrices, $\mathbf{G}_1, \ldots, \mathbf{G}_q$. One may generalize this model by allowing one or more of the variance components to vary from group to group or in accord with some covariates.

The model descriptions above correspond to models known as variance component models (Searle et al., 1992). They do not include covariances between the random effects. Models that include covariances come when the effects within and/or between source of variation are correlated. Therefore, **G** as well as **R** may contain non-zero off-diagonal elements. Laird and Ware (1982) consider the unstructured model for a covariance matrix, i.e. the more general case where all elements of the matrix are allowed to be different. For example, if the genotypes, as in the previous example, were genetically related, a matrix of genetic relationship, **A**, may be used to adjust the variance-covariance matrix of genetic effects. These relationships may be computed from pedigree or molecular based analyses (Falconer, 1989; Bernardo, 1994). Therefore, the matrix of variance-covariances for the **u** vector is $\mathbf{G}=\sigma_u^2 \mathbf{A}$ where elements in **A** are used to represent genetic relatedness between any two genotypes and is expressed as a proportion of the genetic variance. Thus, $\mathbf{A} = \mathbf{I}$ represents the special case of unrelated genotypes. For example in an experiment evaluating half-sib genotypes, the genetic variance is $1/4\ \sigma_a^2$ where σ_a^2 represents additive genetic variance, if the parents themselves are not related (Kang, 1994). So, the parameter function $4\gamma/(1+\gamma)$ represents a narrow-sense heritability. By defining **A** as a matrix with the coefficients of the additive variance for the covariance between genotype *i* and *j* as the *ij*-th element of the matrix, more complex pedigree structures can be considered to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

estimate additive variances. Experimental correlations among observations may be modeled by the off-diagonal elements of **R**. When data are indexed in space, covariances in **R** may reflect correlations due to the spatial unit arrangements

Searle et al. (1992) and Khuri et al. (1998) widely discuss estimation in mixed linear models. A brief discussion is presented here to outline common procedures that will be used to fit plant breeding-orientated mixed models in the subsequent chapters.

Extending the normal equations to allow estimation by generalized least squares (GLS) procedures, Henderson (1975) proposed the mixed model equations (MME). Solving this equation system, estimations of fixed effects and predictors of random effects can be obtained. If **G**, **R**, and **Z**, and hence **V** are known, the generalized least squares solution for β is the best linear unbiased estimators (BLUE), and the solution for the estimation (prediction) of the random effect is the BLUP (Searle et al., 1992). However, in practice, **V** is usually unknown. Therefore, estimation of covariance parameters usually comes prior to the estimation of β and **u**. After obtaining the estimates of **G** and **R**, and **Z** (if it is not known), the fixed and random effects can be estimated by solving the mixed model equations with $\hat{\mathbf{V}}$ replacing **V**. Assuming that the parameterization of the design and variance-covariance matrices is such that the matrices to be inverted are full rank matrices, the mixed model equations may be represented by

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The solutions can be written as

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}y$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(y - \mathbf{X}\hat{\beta}).$$

Under normality, they are equivalent to the maximum likelihood-based solutions (Searle et al., 1992). Kackar and Harville (1984) gave approximations of standard errors of estimators of fixed and random effects in mixed linear models, and showed that GLS solutions were more efficient than corresponding OLS estimators with unbalanced data.

Several authors have discussed why one would be interested in estimating random effect values, i.e. $\hat{\mathbf{u}}$ (Henderson 1975, Harville, 1990, Robinson, 1991). The BLUP of a random effect represents the expected value of the random effect given the observed data. If the joint distribution of **y** and **u** is

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{X}\beta\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{V} & \mathbf{C}\\ \mathbf{C}' & \mathbf{G}\end{bmatrix}\right)$$

where $\mathbf{C} = \mathbf{Cov}(\mathbf{y},\mathbf{u}) = \mathbf{ZG}$. Then the conditional distribution of **u** given **y** is

$$(\mathbf{u} \mid \mathbf{y}) \sim N\left(\mathbf{E}(\mathbf{u}) + \mathbf{C}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta),\ \mathbf{G}-\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}\right).$$

Assuming normality the conditional expectation, $\mathbf{E}(\mathbf{u}|\ \mathbf{y})$, is equal to $\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(y - \mathbf{X}\beta)$ when $\mathbf{E}(\mathbf{u}) = \mathbf{0}$. This is a common expression for the BLUP of the random vector **u**. In reality, the conditional expectation above will be the BLUP of **u** only when **V** is known (Searle et al., 1992). The equation is also a Bayes estimator under a normal *priori*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(Robinson, 1991). In practice, the variance covariance structure of the data is estimated.

The BLUP of a linear combination of fixed and random effects is the linear combination of the BLUE of fixed effects and the BLUP of random effects (Searle et al., 1992). Consider a simple model with one random effect representing genotypic effects and **y** phenotypic data. The prediction equation for genotype *j*, $\varpi_j = \mu + u_j$ is

$$\hat{\varpi}_j = \hat{\mu} + h^2 (y_j - \hat{\mu}).$$

where μ is the population mean, and h^2 is the weighting or shrinkage factor. If **G** and **R** are the traditional structures of the between and within mixed model ANOVA, the elements of $\mathbf{C'V^{-1}}$ are functions of $\sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$ which is the heritability measure associated with **y**. A BLUP is a centered, fixed effects estimate shrunken toward μ, with more shrinkage taking place for smaller values of the estimated variance components in $\mathbf{C'V^{-1}}$, i.e. heritability for this model.

Hitherto, it should be clear that variance components relating to random effects and error terms are needed to obtain estimates of fixed and random effects. Variance component parameter estimates in plant breeding have been typically derived from the expected mean squares of ANOVA tables (Falconer, 1989). This approach, at its best with balanced experiments, is awkward, and at its worst, with unbalanced data, it can seem intractable (Stroup, 1989). ANOVA-based estimators of variance components, which rely on equating mean squares to their expectations and solving for the unknown variance, have nice statistical properties when data are balanced.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

By specifying a Gaussian (normal) distribution for the random effects, the estimation of the unknown parameters is usually obtained using likelihood-based procedures (Hayman, 1960; Harville, 1977). A restricted maximum likelihood method (REML) (Patterson and Thompson, 1971) is usually preferred for estimating the variance components in a mixed model (Searle et al., 1992). Searle (1971) indicated that REML estimates in balanced designs are identical to estimates based on the expected mean squares of ANOVA. For unbalanced data, REML can offer significant advantages over ANOVA-based estimators because REML estimates are unique, non-negative, and have maximum likelihood, large sample statistical properties. The asymptotic standard errors of the estimated variance components can be derived readily as part of the estimation procedure (Searle et al., 1992). In many plant breeding situations, a normal distribution for the data can be realistically assumed, and hence REML approaches are appropriate. Nevertheless, Banks et al. (1985) demonstrated that REML estimates of variance components are robust to violations of this assumption.

The REML procedure of estimating variance components maximizes the residual likelihood function, which is the likelihood function of a set of linear combinations of observed values whose expectations are zero (error contrasts or residuals). Those values are usually obtained by transforming the observations with the transformation matrix $\mathbf{M} = \mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ (Searle et al., 1992). The error contrasts are free of any fixed effects in the model. Thus, the residual likelihood function depends only on the unknown parameters that belong to the variance-covariance structure. The maximization of this function requires numerical procedures. Computation may be extensive with many

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

variance-covariance parameters. Over-parameterized models may be avoided by an appropriate experimental design in relation to the number of parameters to be estimated (Wolfinger and Tobias, 1998).

To do model selection in the mixed model framework, a log likelihood-ratio test criterion can be used with nested models. The procedure demands the evaluation of the restricted log-likelihood (LL_R) for the reduced model (model with smaller number of parameters) and for the full model (model with higher number of parameters).

The test criterion for the likelihood ratio test is,

$$L = -2\{ LL_R \text{ (reduced model)} - LL_R \text{ (full model)}\}.$$

Under normality for the null hypothesis that the reduced model is not different from the full model, the likelihood ratio statistic is distributed as a ^{2} with degrees of freedom equal to the difference in the number of parameters of both models. If the fixed part of the two mixed models under comparison is the same, the test is comparing the covariance structure models. Information criteria such as the Akaike's Information Criterion (AIC) (Sakamoto et al., 1987) are used to compare any set of mixed models. The AIC is the LL_R adjusted for the number of parameters (p) in the model. The adjustment tends to favor parsimonious models. The larger the AIC the more preferable the model (Wolfinger, 1993).

Finally, it is important to note that when using mixed model prediction, optimality properties of the predictors are unknown when the variance parameters are estimated. The typical database structure of a breeding program, i.e., multiple factors with incomplete and unbalanced data, further complicates the analytic evaluation of the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

predictions. From an applied perspective, cross-validation and simulation have been used as validation procedures to assess the accuracy of several types of predictors under particular plant breeding circumstances (Hill and Rosenberg, 1984; Piepho, 1994; Bernardo 1994, Panter and Allen, 1995).

This study compared various mixed models for predicting cross performance, and *per se* performance at early and late stages of a sugarcane-breeding program. Models were assessed for empirical prediction accuracy using cross-validation procedures.

2. REFERENCES

Anderson, R.L., Bancroft, T.A. 1952. Statistical Theory in Research. McGraw-Hill, New York.

Banks, B.D., Mao, I.L., Walter, J.P. 1985. Robustness of the restricted maximum likelihood estimator derived under normality as applied to data with skewed distributions. J. Dairy Sci. 68:1785-1792.

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34:25-30.

Bernardo, R. 1995. Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci. 35:141-147.

Bernardo, R. 1999. Marker-assisted best linear unbiased prediction of single-cross performance. Crop Sci. 1277-1282.

Bjornstad, J.F. 1990. Predictive likelihood: A review. Stat. Sci. 5:242-265.

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

Brown, J., Caligari, P.D.S., Dale, M.F.B., Mackay, G.R., Swan, G.E.L. 1988. The use of cross prediction methods in a practical breeding program. Theor. Appl. Genet. 76:33-38.

Brown, J., Dayle, M.F.B. 1998. Identifying superior parents in a potato breeding program using cross prediction techniques. Euphytica 104:143-149.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Butler, R.W. 1986. Predictive likelihood inference with applications (with discussion). J.Roy. Statist. Soc. Ser. B 48:1-38.

Caligari, P.D.S., Brown, J., Abbott, R.J. 1986. Selection for yields and yield components in the early generations of a potato breeding program. Theor. Appl. Genet. 73:218-222.

Chang, Y.S., Milligan, S.B. 1992. Estimating the potential of sugarcane families to produce elite progeny using univariate cross prediction methods. Theor. Appl. Genet. 84:662-671.

Chang, Y.S. 1996. Assessment of genetic merits for sugarcane parents. Taiwan Sugar Res. Inst. 153:1-9.

Cockerham, C.C. 1963. Estimation of genetic variances. Statistical Genetics and Plant Breeding. Nat. Acad. Sci. Nat. Res. Council Publ. 982, 53-94.

Cox, M.C., Hogarth, D.M. 1993. Progress and changes in the south Queensland variety selection program. Proc. Aust. Soc. Sugar Cane Technol. 15:251-255.

Cox, M.C., McRae, T. A., Bull, J.K., Hogarth, D.M. 1996. Family selection improves the efficiency and effectiveness of a sugarcane improvement program. *In*: Wilson, J.H., Hogarth, D.M., Campbell, J.A., and Graside, A.L. (eds). Sugarcane:Research Towards Efficient and Sustainable Production. CSIRO Division of Tropical Crops and Pastures, Brisbane, 42-43.

Cox, M.C., Stringer, J.K. 1998. Efficacy of early generation selection in a sugarcane improvement program. Proc. Aust. Sugar. Cane Technol. 20:148-153.

Cullis, B.R., Thompson, F.M., Fisher, J.A., Gilmour, A.R., Thompson, R. 1996. The analysis of the NSW wheat variety database. II. Variance component estimation. Theor. Appl. Genet. 92:28-39.

Falconer, D.S.1989. Introduction to quantitative genetics. Ronald Press Co., New York, N.Y. Third edition, Wiley, New York.

Federer, W.T. 1997. Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments. Biometrics

Federer, W.T., Wolfinger, R.D. 1998. SAS code for recovering inter-effect information in experiments with incomplete block and lattice design. Agron. J. 90:545-551.

Gelman, A.J., Carlin, J.B., Stern, H.S., Rubin, D. 1995. Bayesian Data Analysis. Chapman-Hall, New York.

Gogel, B.J., Cullis, B.R., Verbyla, A.P. 1995. REML estimation of multiplicative effects in multi-environment variety trials. Biometrics 51:744-749.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Gopal, J., Gaur, P.C., Rana, M.S. 1992. Early generation selection for agronomic characters in a potato breeding program. Theor. Appl. Genet. 84:709-713.

Harville, D.A. 1976. Extension of the Gauss-Markov theorem to include the estimation of random effects. Ann. Statist. 2:384-395.

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Stat. Assoc. 72:320-340.

Harville, D.A., Callanan, T.P. 1990. Computational aspects of likelihood-based inference for variance components. *In* D. Gianola and K. Hammond (eds.). Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, 136-176.

Harville, D.A.1990 . BLUP Best linear unbiased prediction and beyond. *In* Gianola D., K. Hammond (Eds.). Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, 239-276

Hayman, B.I. 1960. Maximum likelihood estimation of genetic components of variation. Biometrics 16, 369-381.

Henderson, C.R. 1953. Estimation of variance and covariance components. Biometrics 9:226-252.

Henderson, C.R. 1973. Sire evaluation and genetic trends. In W.D.Havey (ed.) Proc. Of Animal Breeding and Genetics Symposium in Honor of J.F.Lush, Virginia Polytech Institute and State University. Am. Soc. Animal Sci. and Dairy Sci. Assoc. Champaign, IL. p. 10-41.

Henderson, C.R. 1974. General flexibility of linear model techniques for sire evaluation. J. Dairy Sci. 57:963.

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

Henderson, C.R. 1990. Statistical methods in animal improvement: Historical Overview. *In* Advances in Statistical Methods for Genetic Improvement of Livestock, New York:Springer-Verlag, 1-14

Hill Jr., R.R., Rosenberg, J.L. 1985. Models for combining data from germplasm evaluation trials. Crop Sci 25:467-470

Hinkley, D.V. 1979. Predictive likelihood. Ann. Statist. 7:718-728.

Hogarth, D.M. 1971. Quantitative inheritance studies in sugarcane. II. Correlation and predicted response to selection. Aus. J. Agric. Res.,22:03-109.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Jenrich, R.L., Schluchter, M.D. 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42:805-820.

Kackar, R.N., Harville, D.A. 1984. Approximations of standard errors of estimators of fixed and random effects in mixed linear models. Comm. Stat., A. Theory and Methods, 10:1249-1261.

Kang, M.S. 1994. Applied Quantitative Genetics. M.S. Kang Publisher, Baton Rouge, LA.

Kang, M.S., Gauch, H.G. (eds). 1996. Genotype-by Environment Interaction. Boca Raton, FL: CRC Press

Khuri, A.I., Mathew, T. Sinha, B.K. 1998. Statistical tests for mixed linear models. Wiley series in Probability and Statistics. John Wiley & Sons, Inc. New York. 352pp.

Laird, N.,Ware, J.H. 1982. Random-effects models for longitudinal data. Biometrics 38:963-974.

LaMotte, L.R. 1973. Quadratic estimation of variance components. Biometrics 29:311-330.

LaMotte, L.R., McWhorter Jr., A., Prasad R.A. 1988. Confidence intervals and tests on variance ratio in random models with two variance components. Comm. Stat. A. Theory and Methods 17:1135-1164.

Liang, K.Y., Zeger, S.L. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13-22.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. 1996. SAS® System for Mixed Models. Cary, N.C.:SAS Institute Inc.

Magari, R., Kang, M.S. 1997. SAS-STABLE: Stability analysis of balanced and unbalanced data. Agron. J. 89:929-932.

McLean, R.A., Sanders, W.L., Stroup, W.W. 1991. A unified approach to mixed linear models. American Statistician 45:54-64.

Milligan, S. B. 1994. Test site allocation within and among selection stages of a sugarcane breeding program. Crop Sci. 34: 1184-1190.

Milligan, S.B., Legendre, B.L. 1991. Development of practical methods for sugarcane cross appraisal. J. Am. Soc. Sugar Cane Technol. 11:59-68.

Milliken, G.A., Johnson, D.E. 1989. Analysis of messy data Volume 2: Nonreplicated experiments. New York: Van Nostrand-Reinold

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Nyquist, W.E. 1991. Estimation of heritability and prediction of selection response in plant populations. Crit. Rev. Plant Sci. 10(3): 235-322.

Oman, S.D. 1991. Multiplicative effects in mixed models analysis of variance. Biometrika 78:729-739

Panter, D.M., Allen, F.L. 1995. Using best linear unbiased predictions to enhance breeding for yield in soybean: I. Choosing parents. Crop Sci. 35:397-405.

Patterson, H.D., Thompsom, R. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58, 545-554.

Piepho, H.P. 1994. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects multiplicative interaction (AMMI) analysis. Theor Appl Genet 89:647-654

Piepho, H.P. 1997. Analyzing genotype-environment data by mixed models with multiplicative effects. Biometrics 53:761-766

Piepho, H.P. 1998a. Stability analysis using the SAS System. Agron. J.

Piepho, H.P. 1998b. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor Appl Genet 97:195-201.

Rao, C.R. 1973. Linear Statistical Inference and Its Applications (second edition) Wiley, New York, NY.

RoBinnson, G.K. 1991. The BLUP is a good thing: the estimation of random effects. Stat. Sci. 6:15-51.

Sakamoto, Y., Ishiguro, M., Kitagawa, G. 1987. Akaike Information Criterion Statistics. KTK Scientific Publisher, Tokyo, Japan.

SAS Institute. 1996. SAS/STAT software: changes and enhancements through release 6.11. SAS Inst., Cary, NC.

Scheffé, H. 1956. Alternative models for the analysis of variance. Ann. Math. Stat. 27:251-271.

Searle, S.R., Casella ,G., McCulloch, C.H. 1992. Variance components. Wiley, New York

Searle, S.R. 1971. Linear models. Wiley, New York

Searle, S.R. 1987. Linear models for unbalanced data. Wiley, New York

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Shafii, B., Price ,W.J. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. Journal Agric., Biol. and Environ Stat. 3:335-345.

Shukla, G.K. 1972. Some statistical aspects of partitioning genotype-environment components of variability. Heredity 29:237-245.

Simmonds, N.W. 1996. Family selection in plant breeding. Euphytica 90:201-208.

Skinner, J.C. 1971. Selection in sugarcane: A review. Proc. Int. Soc. Sugar Cane Technol. 14:149-162.

Stroup, W.W. 1989. Why mixed models?. *In* Applications of Mixed Models in Agriculture and Related Disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge, Louisiana, 104-112.

White, T.L., Hodge, G.R., Delorenzo, M.A. 1986. Best linear prediction of breeding values in forest tree improvement. *In* Workshop of the Genetics and Breeding of Southern Forest Trees, Southern Region Information Exchange Group 40, Gainsville, Fl., 99-122.

Wolfinger, R.D. 1993. Covariance structure selection in general mixed linear models. Comm. Stat. A, Theory and Methods 22:1079-1106.

Wolfinger, R.D., O'Connell, M. 1993. Generalized linear mixed models: a pseudo-likelihood approach. J. Statist. Comput. Simul. 48:233-243.

Wolfinger, R.D., Federer, W.T., Cordero-Brana, O. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. Agron. J. 89:856-859.

Wolfinger, R.D., Tobias R., 1998. Joint estimation of location,dispersion, and random effects in robust design. Technometrics, 40(1):62-71

Zaunbrecher, R.D. 1994. Improving selection procedures in sugarcane using cross appraisal methods. M. Sc. Thesis. Louisiana State Univ., Baton Rouge.

Zeger, S.L., Liang, K.Y., Albert, P.S. 1988. Models for longitudinal data: A generalized estimating equation approach. Biometrics 44:1049-1060.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 2

USING MIXED MODELS TO PREDICT SUGARCANE CROSS PERFORMANCE

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. INTRODUCTION

The initiation of a sugarcane selection cycle starts with the hybridization of parents. Cross appraisal or progeny-tests are often used to focus selection for the best individuals from the best crosses (Hogarth, 1971; Cox et al., 1996; De Sousa-Vieira and Milligan, 1999). The parental information obtained from cross appraisal tests may also be used to predict the best (high mean performance) new hybrids to make.

Commonly, statistical models of progeny test databases adjust data for fixed trial and replication effects and then estimate, from the adjusted data, the untested cross value as a function of the genetic worth of tested crosses (Panter and Allen, 1995). To predict the mean performance of new crosses (crosses that have never been tested before) using progeny test data, the raw means or perhaps rank percentiles of the tested parents are averaged to obtain the mid-parent value (MPV) for a new cross (Caligari and Brown, 1986). Databases developed after some years of progeny tests are commonly incomplete and unbalanced because not all possible crosses between the potential parents are made, only a few parental combinations may be repeated across time, and certain parents are typically used more than others. The irregular data structure creates theoretical and practical concerns about the fixed model approach underlying the MPV prediction (Henderson, 1973; White et al., 1986).

If a narrow inference space is acceptable, an alternative mixed model regarding genetic effects as random and other effects, such as year and location (trial effects), as fixed can be used (McLean, 1991). Several types of best linear unbiased predictors (BLUPs) obtained from the mixed model framework, have been used successfully to help plant breeders choose parents for the best hybrid combinations (Panter and Allen,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1995; Bernardo, 1994, 1995,1996a, 1996b, 1999). Chang and Milligan (1992), Chang (1996), and Cox and Stringer (1998) published BLUP-based analyses related to cross selection in sugarcane. The objective of these analyses was to rank the tested crosses according to the BLUP of their genetic effect (population selection) rather than to predict performance of untested crosses.

When using genetic effect-BLUPs, different random effects and structures for the covariance matrix among those random effects can be postulated. Henderson (1975) described BLUPs of breeding values of potential parents, by using the additive genetic variance relationship among individuals (Henderson, 1976) as the variance-covariance matrix of random genetic effects for each individual. He assumed that the additive genetic variance is the only component of the covariance between observations taken on different individuals.

Models involving female and male random parental effects and genetic covariances among parents have been used successfully to predict single-cross performances of untested hybrids in maize (Bernardo, 1994, 1996b, 1999). Chang (1996) suggested that using the genetic covariances among the parents of the sugarcane crosses to modify the predictions would not be fruitful. He speculated that the highly selected nature of the parents might vitiate the value of such covariances since the genetic covariances estimated from pedigree analysis assume randomly selected parents. Parents used in sugarcane crosses are highly selected.

The offspring used in a progeny test are not selected and hence using the genetic covariances among crosses may enhance the predictive value of BLUPs that incorporate such relationships. Pedigree and/or molecular marker information may be used to set up

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

genetic covariances (based on coancestry) among clones (Bernardo, 1994). Sugarcane molecular information about yield traits is lacking at this time; therefore genetic relationships must be based on the pedigree.

This study compared four predictors for the mean performance of future sugarcane crosses. The predictors were the traditional MPV (fixed model prediction) and three versions of BLUP (mixed model prediction) based on regularly available progeny test information.

2. MATERIALS AND METHODS

2.1. Cross Prediction Models

All cross performance predictors were obtained from models adjusted for fixed trial and replication within trial effects (Panter and Allen, 1995; Table 2.1). The mid-parent BLUP (MP-BLUP) is based on a two-way classification model for the genetic effects involving random "female" and "male" parental effects. I assumed no relationships among the parents. The other two predictors, independent-cross BLUP (IC-BLUP) and related-cross BLUP (RC-BLUP), were based on a one-way classification model involving a random cross effect to model the genetic portion of the response. The difference between these two is that for IC-BLUP the cross effects are assumed to be independent, whereas for RC-BLUP, the additive genetic relationship among crosses is used to set up covariances among the cross effects. By tracking the parent identification of each cross and assuming parents are not related, covariances among crosses within and between trials are simple to obtain based on the cross parentship. I used coancestry coefficients (Falconer, 1989) to establish the additive genetic relationship (covariances) between cross effects when it was required by the predictor equation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2.1. Models used to predict cross performance.

Model†	Effect assumptions	Predictor name
[1] $Y_{ijkl} = \mu + T_k + R_l(T) + F_i + M_j$	all fixed effects	Mid-parent value (MPV)
[2] $Y_{ijkl} = \mu + T_k + R_l(T) + F_i + M_j$	F_i and M_j random else fixed	Mid-parent BLUP (MP-BLUP)
[3] $Y_{ijkl} = \mu + T_k + R_l(T) + C_{ij}$	C_{ij} random independent else fixed	Independent cross BLUP (IC-BLUP)
[4] $Y_{ijkl} = \mu + T_k + R_l(T) + C_{ij}$	C_{ij} random related‡ else fixed	Related cross BLUP (RC-BLUP)

† T_k represents *k*-th trial effect, $R_l(T)$ replication *l* within trial *k* effect, F_i *i*-th female parent effect, M_j *j*-th male parent effect, C_{ij} effect of the combination between female *i* and -male *j*; all model equations include a $N(0, \sigma_e^2)$ random variable as an error term.

‡ Covariances among C_{ij} random effects are obtained from coancestry coefficients.

In matrix notation, the model and derived predictors are specified as:

Model [1] – "Fixed female-male or mid-parent model"

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

where **y** is an *N* vector of observed progeny data; *N* is the total amount of data for a trait; β is the vector of fixed parameters including trial, replications within trials, and female and male parent effects; **X** is incidence matrix relating **y** with β; **e** is an *N* vector of error terms assumed to be normally distributed with zero mean and variance-covariance matrix $\mathbf{R} = \sigma_e^2 \mathbf{I}_N$, where $\mathbf{I}_N$ is a $N \times N$ identity matrix. To obtain the mid-parent values (MPV), i.e., the predictors, female and male parent means, after adjusting by trial and replication effects, were first calculated for each clone used as a parent in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

crosses evaluated in cross appraisal trials. For each potential new cross, corresponding female and male-parent adjusted means were averaged. Note that the parental means were based on progeny data, not *per se* performance data.

Model [2] – "Random female-male model"

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_f\mathbf{a}_f + \mathbf{Z}_m\mathbf{a}_m + \mathbf{e}$$

where β contains the fixed trial and replication within trial effects. The general combining ability (GCA) effects of female parents are in the $f \times 1$ random effect vectors, $\mathbf{a}_f$, where f represents the number of distinct clones used as female parents across trials. The random GCA effects for the male parents are the elements of the $m \times 1$ vector, $\mathbf{a}_m$. It is assumed that $\mathbf{a}_f$ and $\mathbf{a}_m$ were vectors of normal independent random variables with mean zero and variance σ_f^2 and σ_m^2. The performance predictor for an untested cross evaluated under this model, is the mid-parent BLUP (MP-BLUP). The MP-BLUP is the mean of the BLUPs for the female and male effects of the parents involved in the new cross.

Model [3]-"Random independent cross model" is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$, $\mathbf{X}$, β and $\mathbf{e}$ are as before and $\mathbf{u}$ is a c-dimensional vector of random effects representing the genetic effects of c tested crosses; $\mathbf{u}$ is assumed to be normally distributed with zero mean and variance-covariance matrix given by $\mathbf{G} = \sigma_u^2\mathbf{I}$. The performance predictor for potential crosses, y_p, was obtained under the assumption of a normal joint distribution for the random variables in the $c + p$ vector of c tested cross effects and p potential or untested cross effects.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Therefore, $\mathbf{y}_p = \mathbf{CV}^{-1}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta^0)]$, where β^0 is the generalized least squares solution for the vector of fixed effects, and **C** is a $p \times c$ matrix of genetic covariances between the p potential crosses and the c actually tested crosses. Covariances were equated to $2r_{ij}\sigma_a^2$, where r_{ij} is the coancestry coefficient between cross i and j and σ_a^2 was the additive genetic variance. According to the assumptions of unrelated parents, most of the crosses in the data set were either half-sib ($2r_{ij} = ½$), or unrelated ($2r_{ij} = 0$). Thus, by using **C** as the genetic variance-covariance matrix among tested and untested crosses, related crosses contribute to the predicted value for one another. Since cross effects for the tested crosses were treated as independent, the phenotypic variance-covariance matrix for the observed average performance of tested crosses, **V**, contains phenotypic variances equal to $\sigma_a^2 + \sigma_e^2/n_i$ on the main diagonal and zeros on the off-diagonal; n_i is the number of replications for the i-th cross. The cross predictor under this approach was named cross independent-BLUP (IC-BLUP) because of the structure of **G**.

Model [4]-"Random related cross model" was the same as [3] but $\mathbf{G} = \sigma_a^2\mathbf{A}$, where σ_a^2 represents additive genetic variance and **A** is a matrix of $2r_{ij}$ values among tested crosses.

2.2. Data and Validation Procedure

Mean cross performance with regard to cane yield per plant, stalk number per plant, stalk weight, stalk diameter and stalk height was predicted using data from the Louisiana Agricultural Experiment Station (LAES) Sugarcane Variety Development Program's cross appraisal trial data of 719 crosses evaluated between 1992 to 1996 at

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the St. Gabriel Research Station (Bischoff et al., 1994; Bischoff et al. 1995; Bischoff et al., 1996). These trials employed a two replication, randomized complete-block design with approximately 32 progeny from each cross in each 2-row plot. Each row contained approximately 16 plants spaced 41 cm apart within the row and 1.8 m between rows. Not all female parents were combined with each male parent, and only rarely were the same female-male combinations repeated in different years. Between 130 and 200 crosses were tested each year. Cane yield per plant was estimated from stalk counts and estimated stalk weights. Stalk weights were estimated from stalk height and diameter measures from five stalks per row with each stalk from a different plant.

Variance components, σ_e^2 and σ_f^2, σ_m^2, or σ_u^2, were estimated by REML (Searle et al., 1992). For model [2], the female and male parent BLUPs were obtained from the solution of the mixed model equations, after substituting appropriate estimated variance components into a SAS-Proc Mixed (SAS Inst., 1997) program. The covariance coefficients of matrix **C**, for models [3] and [4], as well as those for **A** of model [4] were determined from the parental pedigree of each cross, assuming the parents themselves were unrelated. Estimates of σ_a^2 were obtained using the assumption that $\sigma_f^2 = ¼\,\sigma_a^2$ and $\sigma_m^2 = ¼\,\sigma_a^2$.

The optimality properties for BLUPs and its linear combinations are known when variances are known. In this study, variance component estimates are used, so the analytical properties of the final predictors are unknown. Therefore, I investigated their prediction accuracy when applied to sugarcane data by a "leave-one-trial-out" procedure. The five-year progeny-test database was divided into two data sets, one with

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

four years of trials and the other with the information from the remaining trial. The four-trial predictor data set was used to obtain the cross predictors for the left-out validation trial. The crosses in the left-out trial simulated untested cross performances, since there was no direct information about them in the predictor data set. Predicted values from the four models were obtained for each "untested cross" and compared to mean values of the cross in the validation data set. The validation process was repeated five times, each time using a different (test) year as a validation data set. The squared difference between predicted and observed values was used to approximate prediction error. At each iteration of the validation procedure, counts were made to ascertain how many of the top 50% crosses in the validation set would have been identified by selecting the top 50% of these crosses using the predictions obtained from the prediction database. The average percent of crosses in the top 50% of both groups were expressed as P(50|50). This measure was derived to assess the functional effect of these predictors on selection efficiency. The predictors in the mixed model framework were obtained by a SAS/IML code written to solve genetic mixed models by obtaining parameter estimates from SAS/Proc Mixed. In some cases, they were coupled with calculations of genetic coancestry coefficients from SAS/Proc Inbred (SAS Inst., 1997).

3. RESULTS AND DISCUSSION

In cases of balanced data, known variance parameters and no correlation among genotypes, the MP-BLUP and MPV should lead to the same relative ranking of the genotypes (White et al., 1986). In this study, different results were expected because of the highly unbalanced structure characterizing the progeny test database. For a total of 165 female and 147 male parental clones, only information from 719 parental

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

combinations was available after five years of cross appraisal trials. Almost 90% of the crosses were tested in only one year. Some parents were used in only two hybrid combinations, whereas others were tested in more than 50 crosses. The small number of crosses actually tested is not surprising considering the resource requirements of sugarcane crossing and testing. It draws attention to the importance of cross predictions in sugarcane breeding.

To predict a primary random variable, such as a genetic value, the BLUP first adjusts available data for the fixed effects. After this initial adjustment, the random effects are further adjusted by the fraction of the total variance for which the primary variance accounts, e.g., by heritability (Henderson, 1975). Unbalance is taken into account in the weighting process. For all traits evaluated, the predictor of cross performance with smallest mean square prediction error (MSPE) was based on BLUPs (Table 2.2). The improvement (smaller MSPE) in the prediction accuracy of the BLUP-based predictors with respect to the MPV was consistent across validation data sets. Among the BLUP versions, the MP-BLUP followed the observed data in the trial left out rather closely. The MP-BLUP was the predictor with the smallest MSPE.

The random cross-effect models [3,4] allow one to predict the performance of untested crosses by using the genetic variance-covariance matrix as a link between tested and untested crosses. In this study, the BLUP based on the random cross effect models [3,4] did not perform better than the model constructed from combinations of female and male parent BLUPs [2].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2.2. Prediction accuracy† under four cross performance prediction models for five traits and five test years of sugarcane progeny testing.

Model‡		[1] M + F Fixed	[2] M + F Random	[3] C Random Ind.	[4] C Random Rel.
Predictor		MPV	MP-BLUP	IC-BLUP	RC-BLUP
	Year	----------Root mean square prediction error----------			
Cane yield [kg/plant]	1992	1.09	0.94	0.99	0.97
	1993	0.88	0.69	0.70	0.68
	1994	0.95	0.91	0.95	0.95
	1995	1.04	0.88	0.90	0.87
	1996	1.05	1.00	1.03	1.02
	mean	1.00	0.88	0.91	0.90
Stalk diameter [mm]	1992	1.61	1.38	1.52	1.49
	1993	1.11	1.03	1.24	1.24
	1994	1.26	1.17	1.31	1.33
	1995	1.61	1.24	1.51	1.50
	1996	1.54	1.38	1.65	1.65
	mean	1.43	1.24	1.45	1.44
Stalk height [cm]	1992	17.51	15.58	16.32	16.14
	1993	13.19	12.29	13.16	13.14
	1994	12.06	9.76	11.55	11.57
	1995	17.08	16.49	17.00	16.50
	1996	13.25	11.69	12.36	11.97
	mean	14.62	13.16	14.08	13.86
Stalk number [plant^{-1}]	1992	1.34	0.98	1.03	0.99
	1993	1.30	0.89	0.89	0.85
	1994	1.13	1.16	1.11	1.12
	1995	1.39	1.19	1.28	1.23
	1996	1.96	1.81	1.84	1.81
	mean	1.42	1.21	1.23	1.22
Stalk weight [kg x 10^2]	1992	11.40	10.54	11.02	10.59
	1993	8.42	6.27	7.25	7.13
	1994	10.32	9.03	10.32	9.73
	1995	11.59	7.91	10.01	8.98
	1996	9.56	8.09	9.53	9.46
	mean	10.26	8.37	9.63	9.18

† Square root of the mean square prediction error (difference between predictor and target values for each cross obtained by an iterative "leave-one year-out" validation procedure).

‡ All models take the form: $Y_{ij} = \mu + T + R(T) +$ [as indicated in the column heading] $+ \varepsilon$. Notation for model effects is T - trial, R- replication, C - cross, F - female, M – male, $\varepsilon \sim N(0, \sigma_e^2)$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The one-way (cross) fixed model was not included in the study because it can not be used to make inference about new crosses. It only offers information about crosses already made.

Possible factors for the behavior of BLUPs based on cross effects [3,4] with regard to MP-BLUP [2] might be related to an insufficient number of related crosses per cross, equal weighting of female and male parent related crosses, and high dominance variance. The median for the number of relatives (usually half-sibs) per cross in the database was 34, the first quartile 18, and the third quartile 50; in other words, 50% of the crosses had less than 34 related crosses in the database. Some of them could be related through the female parent and others through the male parent, but all of them were equally weighted in the cross-effects-based BLUPs [3,4]. On the contrary, under model [2], BLUPs of female effects are estimated using the estimation of σ_f^2 and BLUPs of male effects the estimation of σ_m^2.

Even under a half-sib structure, both variance components should theoretically lead to the same additive genetic variance among crosses. In this study there were consistent differences between both variance component estimates for all the traits (Table 2.3). The control of experimental errors related to identification of female and male parents is different when crossing sugarcane at LAES. Sugarcane crosses in the LAES, as in many sugarcane breeding programs, are made in cubicles with a designated "male" parent tassel suspended above several designated "female" parent tassels. The functional sex of the clone is usually assessed as a function of the amount of pollen the clone produces. In many cases, the functional female is not completely male sterile and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

hence could, and probably does, perform some cross-pollination of other females in the same cubicle. Thus, the female parent is known with certainty and the male parent is known with less certainty. The lack of male parent certainty may explain the trend of larger σ_f^2 than σ_m^2 for cane yield, stalk weight and stalk height, but it does not explain the tendency for a larger σ_m^2 being larger than σ_f^2 for stalk diameter and stalk number. Perhaps the differences are not real or perhaps the sampled populations were just different in the tested years. Parental genotypes do change somewhat each year.

Table 2.3. Female and male variance component estimates for five traits and test years of sugarcane progeny testing.

Test year	Variance component	Cane yield	Stalk weight	Stalk height	Stalk diameter	Stalk number
		$(kg/plant)^2$	$(kg \times 10^{-2})^2$	cm^2	mm^2	$(no/plant)^2$
1992	σ_f^2	0.114	0.221	51.74	0.554	0.126
	σ_m^2	0.030	0.131	16.00	0.720	0.123
1993	σ_f^2	0.090	0.186	56.02	0.423	0.050
	σ_m^2	0.052	0.106	22.49	0.493	0.101
1994	σ_f^2	0.114	0.181	63.43	0.390	0.031
	σ_m^2	0.065	0.107	18.72	0.429	0.092
1995	σ_f^2	0.128	0.208	37.77	0.484	0.064
	σ_m^2	0.103	0.106	21.84	0.605	0.108
1996	σ_f^2	0.113	0.186	55.90	0.359	0.032
	σ_m^2	0.035	0.126	21.21	0.477	0.090
Mean residual variance	σ_e^2	1.208 ±0.043	0.958 ±0.034	222.40 ±4.602	1.920 ±0.061	1.184 ±0.059

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The **G** and **C** matrices used for the BLUPs based on cross effects are narrow-sense genetic covariance matrices since they are based on additive variance coefficients. Models introducing dominance variance and specific combining ability may improve predictions. Previous research identified significant dominance effects for yield components in sugarcane (Milligan, 1988). The accuracy of the prediction for the different traits indicated better predictions for stalk diameter, stalk height, and stalk weight than for cane yield and stalk number. This probably reflects the relative heritability of these traits (Milligan, 1988).

The rank correlations between predicted and observed performances of untested crosses ranged from non-significant values for stalk number to an average correlation of 0.52 (P=0.001) for stalk diameter when using MP-BLUP as the predictor (Table 2.4).

Table 2.4. Mean rank correlation between predicted and observed cross values†.

Trait,	Predictor			
	MPV	MP-BLUP	IC-BLUP	RC-BLUP
	---------- mean rank correlation ----------			
Cane yield	0.23	0.35	0.29	0.34
Stalk number	ns	ns	ns	ns
Stalk weight	0.34	0.46	0.35	0.38
Stalk diameter	0.34	0.52	0.37	0.42
Stalk height	0.35	0.48	0.39	0.39

† Mean of significant correlations. For MPV and MP-BLUP values are average of five years. For IC-BLUP and RC-BLUP, values represent 2average across three (significant) out of a total of five test years; ns=no significant correlations (P>0.05).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The average correlations of MP-BLUPs and observed values were also close to 0.50 for stalk height and stalk weight. Cane yield showed smaller correlation coefficients than the other traits did (except stalk number). Correlation coefficients based on MP-BLUPs were consistently better than those obtained from the RC-BLUP, IC-BLUP or MPV. Correlations tended to be higher or more likely significant when the ratio of the genetic to residual variance among untested crosses was higher (Table 2.5). The maximum expected correlation between the predicted and the observed value is not unity but it depends on the heritability of the trait (Bernardo, 1999). This is because we are correlating predicted genotype with phenotypic values. Heritability for cane yield is not superior to 0.30 (Milligan et al., 1990), thus the correlation between genotype and phenotype, $(0.30)^{1/2}= 0.55$, is the upper bound for an observed correlation.

Table 2.5. Rank correlations between predicted and observed cross performances under four cross prediction models for cane yield in five years of sugarcane progeny testing.

Year	σ_g^2/σ_e^2 ‡	Model and predictor†			
		[1] M + F Fixed MPV	[2] M + F Random MP-BLUP	[3] C Random Ind. IC-BLUP	[4] C Random Rel. RC-BLUP
		---------- Rank Correlation----------			
1992	0.35/0.86	0.21	0.32	0.27	0.30
1993	0.23/0.70	0.19	0.30	0.30	0.32
1994	0.14/1.71	0.26	0.36	ns	ns
1995	0.23/1.07	0.24	0.38	0.29	0.39
1996	0.23/1.56	0.24	0.40	ns	ns

† All models take the form: $Y_{ij} = \mu + T + R(T) +$ [column heading] $+ \varepsilon$. Notation for model effects is T - trial, R- replication, C - cross, F - female, M – male, $\varepsilon \sim N(0, \sigma_e^2)$.

‡ Ratios of REML estimators of broad-sense genetic variance and residual variance in the test year left out for validation purposes; ns - no significant correlation, other correlation are significant at $\alpha = 0.05$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bernardo (1992) indicated that a correlation between predicted and true genetic value around 0.60 would allow a breeder to select the top 20% crosses while maintaining at least an 80% chance of retaining the best hybrid in the selected group. Using LAES data, correlations around 0.50 were obtained for stalk diameter, stalk weight, and stalk height, but correlations for cane yield were smaller than 0.40 and several non-significant correlations were detected for stalk number (Table 2.4). Thus, ranking of potential crosses should be better when based on stalk diameter and height.

Modifications to the current testing methodology that should enhance the quality of the test data have been initiated. DeSousa-Vieira and Milligan (1999) demonstrated that increasing the intra-row plant spacing would significantly increase the genetic variance and functional heritability of cane yield and stalk number. A wider intra-row plant spacing (about 60 cm), is now used by the LAES in its progeny testing program than that previously used to generate the data in this study. An additional expected improvement in the progeny testing methodology is the implementation of weighing the entire plot as opposed to estimating cane yield from stalk counts and an estimated stalk weight.

The MP-BLUP [2] improved selection efficiency more than the cross-based BLUPs [3,4] compared to the traditional MPV [1] (Table 2.6). Differences of 10 to 20%, for the percent of top crosses identified when selecting 50% of the potential crosses based on the predictor values, were observed between MP-BLUP and MPV, with consistently higher percentages for MP-BLUP. Identification of top 50% crosses varied from 54% to 72%, depending on the trait and model. They ranged from 54% to 62% of the top

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

half of the crosses identified when using the MPV, versus 59% to 72% for the MP-BLUP, which represents a substantial improvement in selection efficiency.

Table 2.6. Percent of top 50% crosses retained in the validation data set also in the top 50% of the predicted cross performances for four prediction models.

Trait	Model and Predictor†			
	[1] M + F Fixed MPV	[2] M + F Random MP-BLUP	[3] C Random Ind. IC-BLUP	[4] C Random Rel. RC-BLUP
	---------- ---- % ----------			
Cane yield	56	68	60	61
Stalk diameter	60	72	65	69
Stalk height	61	70	61	60
Stalk number	54	59	59	60
Stalk weight	62	71	62	63

† All models take the form: $Y_{ij} = \mu + T + R(T) +$ [as indicated in the column heading] $+ \varepsilon$. Notation for model effects is T - trial, R- replication, C - cross, F - female, M – male, $\varepsilon \sim N(0, \sigma_e^2)$.

Despite predictors from model [3] and [4] showing smaller MSPE than the MPV (Table 2.2), there was not a substantial improvement over the fixed MPV [1] regarding correlations and P(50|50) (Table 2.4 and 2.6). The preference of BLUPs based on tested cross effects [3,4] over MP-BLUP is doubtful for sugarcane. Cox and Stringer (1998) worked with mid-parent BLUPs in sugarcane. They estimated BLUPs based on a complex trait, Net Merit Grade, for parents of families harvested from 1993 to 1995 at the core breeding program in Australia. BLUP estimates were correlated with clonal performance of those families at stage 1 (first clonal stage) in 1994, 1995, and 1996.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Correlations were 0.60-0.65. It is important to note that they evaluated progeny performance using whole plot weights in a way similar to that now employed in the LAES. Their results indicated that MP-BLUP is also advantageous for selecting families to advance genotypes in early generations.

4. CONCLUSIONS

Progeny testing has proved to be effective and cost efficient for sugarcane breeding because it improves efficiency of early generation selection. It can also be exploited to generate BLUPs of untested crosses. Predictors based on progeny tests can be obtained after one year of testing of new parents, whereas parental selection based only on clonal *per se* information requires several years of testing and probably is more biased by the presence of non-additive genetic effects. No additional field experiments are required to calculate cross performance predictors in sugarcane breeding programs involving progeny tests.

Cross prediction mixed models or BLUP-based predictions consistently improved over the fixed MPV model the accuracy of the predicted performance of crosses that have never been tested. A 2-way random model with female and male effects performed better than a one-way cross effect model with covariances adjusted by additive genetic relationship among crosses in sugarcane. The results suggest that the MP-BLUP obtained from the databases of regular sugarcane progeny tests would facilitate the identification of material to be crossed. The accuracy of predictions might be increased by maximizing experimental efficiency of cross appraisal tests, i.e. more

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

replications, better procedures for recording data and combining plant cane with first ratoon data.

5. REFERENCES

Bernardo, R. 1992. Retention of genetically superior lines during early-generation testcrossing of maize. Crop Sci. 32:933-927.

Bernardo, R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34:25-30.

Bernardo, R. 1995. Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci. 35:141-147.

Bernardo, R. 1996a. Best linear unbiased prediction of maize single-cross performance. Crop Sci. 36:50-56.

Bernardo, R. 1996b. Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci. 36:872-876.

Bernardo, R. 1999. Marker-assisted best linear unbiased prediction of single-cross performance. Crop Sci. 1277-1282.

Bischoff, K.P., Milligan S.B., Rodriguez, P.H., Zaunbrecher, R.D., Quebedeaux, K.L., Martin, F.A. 1994. Selections, advancements, and assignments of the Louisiana. "L", sugarcane variety development program for the year 1994. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 23-76.

Bischoff, K.P., Milligan S.B., Rodriguez, P.H, Quebedeaux, K.L., Martin, F.A. 1995. Selections, advancements, and assignments of the Louisiana, L, sugarcane variety development program for the year 1995. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 21-40.

Bischoff, K.P., Milligan S.B., Gravois, K.A., Quebedeaux, K.L., Hawkins, G. L. Martin, F.A. 1996. Selections, advancements, and assignments of the Louisiana. "L", sugarcane variety development program for the year 1996. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 31-54.

Caligari, P.D., Brown, J.,Abbott, R.J. 1986. Selection for yield and yield components in the early generations of a potato breeding program. Theor Appl Genet 73:218-222.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chang, Y.S., Milligan S.B. 1992. Estimating the potential of sugarcane families to produce elite progeny using univariate cross prediction methods. Theor. Appl. Genet. 84:662-671.

Chang, Y.S. 1996. Assessment of genetic merits for sugarcane parents. Taiwan Sugar Res. Inst. 153:1-9.

Cox, M.C., McRae, T. A., Bull, J.K., Hogarth, D.M. 1996. Family selection improves the efficiency and effectiveness of a sugarcane improvement program. *In*: Wilson, J.H., Hogarth, D.M., Campbell, J.A., and Graside, A.L. (eds). Sugarcane:Research Towards Efficient and Sustainable Production. CSIRO Division of Tropical Crops and Pastures, Brisbane, 42-43.

Cox, M.C., Stringer, J.K. 1998. Efficacy of early generation selection in a sugarcane improvement program. Proc. Aust. Sugar. Cane Technol. 20:148-153.

DeSousa-Vieira, O., Milligan, S.B. 1999. Intrarow plant spacing and family x environment interaction effects on sugarcane family evaluation. Crop Sci. 39:358-364.

Falconer, D.S. 1989. Introduction to Quantitative Genetics. Ronald Press Co., New York, NY Third edition, Wiley, New York.

Harville, D.A.1990. BLUP best linear unbiased prediction and beyond. *In* Gianola D., K. Hammond (Eds.). Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, 239-276.

Henderson, C.R. 1973. Sire evaluation and genetic trends. In W.D.Havey (ed.) Proc. Of Animal Breeding and Genetics Symposium in Honor of J.L.Lush, Virginia Polytech Institute and State University. Am. Soc. Animal Sci. and Dairy Sci. Assoc. Champaign, IL. p. 10-41.

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

Hogarth, D.M. 1971. Quantitative inheritance studies in sugarcane. II. Correlation and predicted response to selection. Aust. J. Agric. Res.22:03-109.

Littell, R.C., Milliken, G.A., Stroup ,W.W., Wolfinger, R.D. 1996. SAS® System for Mixed Models. Cary, N.C.:SAS Institute Inc.

McLean, R.A., Sanders, W.L., Stroup, W.W. 1991. A unified approach to mixed linear models. American Statistician 45:54-64.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Milligan, S.B., Gravois, K.A., Bischoff, K.P., Martin, F.A. 1990. Crop effects on broad-sense heritabilities and genetic variances of sugarcane yield components. Crop Sci. 30: 344-349.

Milligan, S.B. 1994. Test site allocation within and among selection stages of a sugarcane breeding program. Crop Sci. 34: 1184-1190.

Panter, D.M., Allen, F.L. 1995. Using best linear unbiased predictions to enhance breeding for yield in soybean: I. Choosing parents. Crop Sci. 35:397-405.

RoBinnson, G.K. 1991. The BLUP is a good thing: The estimation of random effects. Stat. Sci. 6:15-51.

SAS Institute. 1997. SAS/STAT software: changes and enhancements through release 6.12. SAS Inst., Cary, NC.

Searle, S.R., Casella,G., McCulloch, C.H. 1992. Variance components. Wiley, New York

White, T.L., Hodge, G.R., Delorenzo, M.A. 1986. Best linear prediction of breeding values in forest tree improvement. *In* Workshop of the Genetics and Breeding of Southern Forest Trees, Southern Region Information Exchange Group 40, Gainesville, Fl., 99-122.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 3

COMBINING DATA ACROSS TESTS TO PREDICT FUTURE *PER SE* PERFORMANCE IN A SUGARCANE BREEDING PROGRAM

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. INTRODUCTION

Breeders routinely select genotypes or lines in a series of selection stages. The earliest stages after the initial hybridization are typically unreplicated tests in which large numbers of genotypes are screened. As the material advances through stages, the screened number of genotypes or lines drastically shrinks and the extent of testing (replications, plot size, number of locations and amount of data) proportionally increases. Selection and advancement decisions may be based upon several tests. These tests are often carried out in different time periods and vary in experimental dimensions such as the number of entries, plot size, number of replications, genetic level of elitism, and experimental precision. The breeder must computationally, or at least mentally, combine the data to make decisions. He will typically weigh the relative precision of the various experiments and will commonly scale the results in some manner to make the numbers comparable. Therefore, many of the challenges in combining information to improve genotype prediction reduce to matters of scaling and weighting for relative confidence in the data.

Weighted and unweighted analysis of variance provides a useful statistical technique for combining data to analyze treatment (genotype) differences (Milliken and Johnson, 1989). The traditional ANOVA under a fixed model approach assumes independent observations (Searle, 1987). This assumption limits the quantity of data gleaned from breeding databases. Least squares means analysis obtained under a fixed linear model often ignores the underlying correlation structure in the data (Latour and Littell, 1996). A mixed

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

linear model, however, allows an analytical approach to account for genetically and/or experimentally correlated data (Stroup, 1989; Searle et al., 1992). The random nature of genotype effects in early testing may be taken into account in a mixed model approach. Other variables, besides genotype effects, such as blocks, and trial effects, might qualify for consideration as random effects (Wolfinger et al., 1997). Mixed models may account, through the variance-covariance structure, for heterogeneity of standard errors and genetic as well as non-genetic correlations.

Given the typical size of plant breeding data sets, estimation procedures under a regular mixed or a random model (Searle et al., 1992) that involves replications may require extensive computer time and memory (Piepho, 1998). A practical solution is to work with genotype-test means. However, with incomplete designs and heteroscedastic data, the analysis may not be valid (Piepho, 1998). An estimation procedure called weighted two-stage analysis has been applied successfully to sort out the challenges associated with fitting mixed models in the analysis of cultivar trials (Cullis et al., 1996a,b; Frenshman et al., 1997). The main idea behind this procedure is first to estimate variance components for each test and then work with a means model, using the estimated variances as known parameters to weight observations. Mixed models further use best linear unbiased predictors (BLUPs) to estimate random effects (Henderson, 1975; Harville, 1990). Thus, performance predictions for genotypes will be expressed by a BLUP instead of a mean, as is common in the fixed model approach (White et al., 1986; Panter and Allen,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1995). Various versions of BLUPs might be defined depending on the underlying mixed model and goals of the prediction.

Many breeding programs worldwide use check cultivars to facilitate combining information from different sources (Yates and Cochran, 1938; Cochran, 1954; McIntosh, 1983; Hill and Rosenberg, 1985). For example, the Louisiana Sugarcane Variety Development Program (LSVDP) expresses experimental genotype yields as a percentage of each check in the same replication in a given trial. The overall mean of these percent-of-the-check values is used in selection and advancement decisions (Milligan et al., 1994). The use of check varieties is not free from concerns. The checks may be unstable in performance and usually change with time, as different and more relevant varieties become important reference cultivars. Furthermore, check cultivars may be quite different from the experimental population under investigation. Treating them as a member of the experimental genotype population may bias genetic variance component estimates and hence predictors.

This study investigated three mixed models, involving three versions of BLUPs estimated under different strategies, least squares genotype means under a fixed model, and four check-based methods for combining information. The goal was to compare strategies to predict future genotype performance.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. MATERIALS AND METHODS

2.1. Models for Combining Early Selection Stage Data

All models combined trial information into a single estimate to predict the genotype performance in future trials. Besides the overall arithmetic mean, two types of predictors were analyzed: check-based predictors and linear model-based predictors.

2.1.1. Check based-predictors

I examined four genotype performance predictors that used checks in their derivation (Table 3.1). The first one was the "average percent of the check" (APCH) where experimental genotype yields were expressed as a percentage of different commercial check values in the same replication. The predictor, $APCH_j$ [1], is the average of these values across checks, trials and replications for genotype *j*.

The second check-based predictor expresses the genotype values in a trial as a percentage of the mean of all the checks in that trial. The predictor, $PACH_j$ [2], is the average of this value over all tests that contain the genotype *j*.

The third check-based predictor is the average difference, AD_j [3], between each experimental genotype *j* in rep *k* and trial *i*, and the mean of all the checks in that trial.

The predictor, AD_j, is the average across all trials and replications for genotype *j*. The fourth predictor is the standardized experimental genotype value within each trial and replication (SP_{ijk}) using mean and standard deviation of the checks in the trial. The SP_j value for a given genotype *j*, is the average across all trials and replications.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3.1 Predictors of *per se* genotype performance.

Predictor	Formula/Model†	Assumptions and comments
[1] Average Percent of Checks (APCH)	$APCH_j = \sum_{i=1}^{I} (\sum_{k=1,m=1}^{KM} (100\, y_{ijk} / C_{ikm})/MK)/I$	Check based
[2] Percent of Average Check (PACH)	$PACH_j = \sum_{i=1,k=1}^{IK} (100\, y_{ijk} / \bar{C}_{i..})K$	Check based
[3] Average Difference (AD)	$AD_j = \sum_{i=1,k=1}^{IK} (y_{ijk} - C_{ikm})/IK$	Check based
[4] Standardized Performance (SP)	$SP_j = \sum_{i=1,k=1}^{IK} (y_{ijk} - \bar{C}_{i..})/\sigma_{ci})/IK$	Check based
[5] Least squares Means (LSM)	$y_{ijk} = \mu + T_i + R(T)_{ik} + G_j + e_{ijk}$	All effects fixed except e_{ijk}
[6] BLUPs of genotype effects (BGa)	$y_{ijk} = \mu + T_i + R(T)_{ik} + G_j + e_{ijk}$	G_j random, T_i and $R(T)_{ik}$ fixed
[7] BLUPs of genotype effects (BGb)	$y_{ijk} = \mu + T_i + R(T)_{ik} + C_m + G_j + e_{ijk}$	Checks (C_m) separated; G_j random, else fixed
[8] BLUPs of genotype effects (BGTa)	$y_{ijk} = \mu + T_i + R(T)_{ik} + G_j + e_{ijk}$	All effects random
[9] BLUPs of genotype effects (BGTb)	$y_{ijk} = \mu + T_i + R(T)_{ik} + C_m + G_j + e_{ijk}$	Checks (C_m) separated; all effects random except checks
[10] BLUPs of genotype effects (BGTIa)	$y_{ij} = \mu + T_i + G_j + GT_{ij}$	Trial means used; all effects random; unweighted
[11] BLUPs of genotype effects (BGTIb)	$y_{ij} = \mu + T_i + G_j + GT_{ij}$	Trial means used; all effects random; weighted

† C_{mk} trait value for check m (m=1,...,M) at trial i, replication k; σ_{ci} std. deviation of check values at trial i; y_{ijk}, (i = 1, ..., I; j = 1, ..., J; k =1,..., K) trait value for j-th genotype at trial i; general mean; G_j genotype effect; T_i environment effect; $R(T)_{ik}$ block within trial; GT_{ij} interaction.

2.1.2. Two-way linear model-based predictors

Another set of predictors was obtained using the regular linear model:

$$y_{ijk} = \mu + T_i + R(T)_{ik} + G_j + GT_{ij} + e_{ijk}$$

where y_{ijk} ($i = 1, ..., I; j = 1, ..., J; k = 1, ..., K$) was the trait value of the replicate k, genotype j and trial i. T_i was trial i effect, $R(T)_{ik}$ was replication k in trial i effect, G_j was genotype j effect, GT_{ij} was the genotype by trial interaction effect, and e_{ijk} was the error term associated with y_{ijk}. In a fixed linear model all model components except the error terms are considered fixed values. Least squares means (LSM) [5] of genotype levels across a given set of trials for a model without GT interaction (interaction terms are non-estimable) were used as genotype performance predictors (Table 3.1). Six versions of BLUPs of genotype performance were obtained using different mixed models. The models varied in their inclusion of $R(T)_{ik}$ or GT_{ij} effects, whether or not check genotype (C_m m = 1, ..., M) were considered fixed or random, and whether or not T_i and $R(T)_{ik}$ were considered random or fixed. All models considered genotype effects as random.

Predictors BGa [6] and BGb [7] were BLUPs from a model that did not include GT effects and considered trial and rep(trial) effects as fixed. Predictor BGa included the checks among the genotype random effect, whereas predictor BGb considered the check effects as fixed and separated them from the experimental genotype effects in the model. Predictors BGTa [8] and BGTb [9] BLUPs were generated using the same models as used for [6] and [7]. However, in addition to genotype effects, trial and replication-within-trial effects were considered random. Predictor BGTa grouped checks with experimental genotypes (all effects random). Predictor BGTb separated checks from the experimental genotypes by considering check effects as fixed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Predictors BGTIa [10] and BGTIb [11] modeled trial means instead of plot values and included the GT. Predictor BGTIa was unweighted for trial residual variances. BGTIb weighted for trial residual variances using r_i/ s^2_i as weights where r_i is the number of replications in trial *i* and s^2_i was the error mean square in the same trial obtained from a previous ANOVA for each trial. Working with trial means instead of individual plot data made the incorporation of the GE effects computationally feasible. The predictor value of each method was expressed in the original trait value range (Table 3.2).

Table 3.2. Predictor conversion to trait unit values.

Predictor	Conversion†
---------- check- based ----------	
[1] Average Percent of Checks (APCH)	$y_j = (APCH_j)(\mu_c)/100$
[2] Percent of Average Check (PACH)	$y_j = (PACH_j)(\mu_c)/100$
[3] Average Difference (AD)	$y_j = (AD_j) + \mu_c$
[4] Standardized Performance (SP)	$y_j = (SP_j)\,\sigma_c + \mu_c$
---------- fixed model ----------	
[5] Least squares Mean genotype effects (LSM)	$y_j = LSM_j$
---------- mixed model ----------	
[6] BLUPs of genotype effects (BGa)	$y_j = BGa_j + \mu^\circ$
[7] BLUPs of genotype effects (BGb)	$y_j = BGb_j + \mu^\circ$
[8] BLUPs of genotype effects (BGTa)	$y_j = BGTa_j + \mu^\circ$
[9] BLUPs of genotype effects (BGTb)	$y_j = BGTb_j + \mu^\circ$
[10] BLUPs of genotype effects (BGTIa)	$y_j = BGTIa_j + \mu^\circ$
[11] BLUPs of genotype effects (BGTIb)	$y_j = BGTIb_j + \mu^\circ$

†σ_c and μ_c equaled the across trials mean and standard deviation for the checks; μ° is the generalized least squares estimator of the overall mean.

For all mixed models, the random effects were considered independent and normally distributed random variables. I assumed that number of replications per

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

genotype might be different within and between trials (unbalanced data) and that not all genotypes were evaluated in all trials in the database (incomplete data). The variance components were estimated by REML using a SAS code based on Proc Mixed/SAS (SAS Inst., 1997) (Appendix A). In addition to model differences regarding the characteristics mentioned, the predictors for genotype performance under the mixed models were empirical BLUPs of genotype effects. The adjective "empirical" is appropriate since the variance components used in the calculation of BLUPs were estimate

2.2. Data and Validation Procedure

The LSVDP database of early yield trials between 1988-1995 was used to compare different models. Personnel in the LSVDP select advance and plant genotypes (clones) among nine clonal stages in sequential years (Table 3.3; Milligan, 1994). Replicated testing begins in the third clonal stage (Increase stage) and multi-location testing in the fourth clonal stage (Nursery stage). The first two clonal stages after crossing are unreplicated due to the high number of plots established in these stages (about 3000 reduced to 1000). About 50 to 70 clones are replicated at three locations in the fourth clonal stage (Nursery). A new series is initiated every year and plots are harvested once a year for three years (plantcane, first ratoon and second ratoon crops).

Only plantcane data were considered in this study. The commercial sugarcane varieties, CP65-357 (Breaux et al., 1974), CP70-321 (Fanguy et al., 1979), CP72-370 (Fanguy and Breaux, 1981), CP74-383 (Fanguy et al., 1983), and LCP82-089 (Martin et al., 1992) were used as checks. A subset of 35 genotypes, involving different

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

assignment series and including the 5 checks, was chosen as the set of genotypes for which performance would be predicted.

Varieties from different assignment series are rarely tested in the same trial. All trials involving these genotypes and all the genotypes included in those trials made the "working data set". Variety data accumulates each year (Table 3.3).

Table 3.3. Stages, dimensions and plantcane yield information of the Louisiana Sugarcane Variety Development Program.

		Dimension				Plots Harvested†	
Year	Stage	Entries	Area	Loc	Reps	Annual	Total
		No.	m^2	----------No.----------			
1	Crossing						
2	Seedling	50,000	0.8	1	1	0	0
3	1st clonal	3,000	3.3	1	1	0	0
4	2nd clonal	900	23.8	1	1	0	0
5	Increase	300	23.8	1	2	1	1
6	Nursery	70	23.8	3	2	3	4
7	Nurs./Infield	30	23.8/71.3	3	2	9	13
8	Introduction‡	10	--		1	18	44
9	Outfield	8	53.5	10	3	16	60
10	Outfield	3	53.5	10	3	40	100
11	Outfield	2	53.5	10	3	60	160
12	Outfield	1	53.5	10	3	90	250
13	Release	1					

† Actual amount of data is typically less because some tests may not be planted (Infield) or something prevents harvest.

‡ No new data collected; seed-cane increase only.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

To assess model prediction accuracy, predictors were derived from earlier years in the selection process (calibration data set) to predict yields in more advanced stages of the program (validation data set). The comparison used data from the 2nd, 3rd, and 4th clonal stages (2nd clonal, Increase, and Nursery trials from year 6) as the calibration data set to predict yields in the 5th clonal stage (Nursery and Infield data from year 7). The second clonal stage, an unreplicated stage, was treated as a third replication of the 3rd clonal stage (Increase stage) that included cultivars of the same assignment series.

Prediction errors were obtained by calculating the difference between the predicted value and the value observed in the validation data set (5th clonal stage). The square root of the average of square differences (RMSPE) was reported as estimations of prediction errors. The procedure was repeated with ten sets of 35 genotypes to calculate mean prediction accuracy. The response variables (traits) predicted were cane yield (Mg ha^{-1}), stalk weight (g), sucrose content (g sucrose kg^{-1}cane) and stalk number (no. m^{-2}). Rank correlations between the predictors based on the 2nd through 4th clonal stage and the mean genotypic values in the 5th clonal stage (validation data set) were also calculated.

3. RESULTS AND DISCUSSION

Prediction errors for all predictors ([1] to [11]) were smaller than those for the raw mean for all traits (Table 3.4). Other than the raw mean, the best and worst predictors varied by trait. The lowest prediction errors were obtained by BGTa [8] for cane yield, stalk weight and sucrose content. This mixed model-based method considered all effects random. The best method for stalk number was the standardized prediction method [4], which used the check trial mean and standard deviation to standardize the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

experimental genotype value. Of the non-mixed model-based methods ([1] to [5]), the SP [4] was the most effective predictor. This result also indicated the importance of adjusting genotype data not only for average trial yield but also for intra-trial variability.

Table 3.4. Prediction accuracy for twelve methods of combining early selection stage data to predict genotype performance in a more advanced stage of a sugarcane breeding program for four traits.

Predictor	Cane yield	Stalk weight	Sugar content	Stalk number
	Mg ha^{-1}	g	g kg^{-1}	no. m^{-2}
	----------Root mean square prediction error----------			
[1] APCH	19.1	57.7	6.74	0.843
[2] PACH	17.2	54.5	6.77	0.802
[3] AD	17.9	61.3	6.17	0.788
[4] PS	15.3	51.3	6.14	0.777
[5] LSM	18.5	64.9	7.28	0.846
[6] BGa	16.6	60.4	5.71	0.827
[7] BGb	16.8	62.2	5.64	0.841
[8] BGTa	14.9	48.6	5.42	0.836
[9] BGTb	15.0	48.6	5.43	0.838
[10] BGTIma	15.7	58.1	6.10	0.856
[11] BGTImb	15.0	53.1	5.80	0.853
[12] Raw mean	26.2	130.8	14.96	1.015
	---------- Trait mean ----------			
Mean	72.8	1117	128.7	8.013

† Square root of the mean square prediction error (difference between predictor and target values for each genotype obtained by an iterative validation procedure).

In general, separating the checks from the experimental genotypes and then considering them as fixed did not improve the predictive value of the mixed model-based predictors. Even with checks that might have high effect values, the number of check varieties is small compared with the set of varieties that participate in the calculation of genotypic variances.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Predictors with the lowest prediction error generally produced higher rank correlations between the calibration data set (2nd, 3rd and 4th clonal stage) and the validation data set (5th clonal stage) (Table 3.5). There were some slight exceptions to this observation in that the BGTImb predictor (weighted mean based predictor [11]) produced somewhat higher correlations than one would expect from the prediction errors. This is a mean-based predictor.

Table 3.5. Rank correlations between predictor based on 2nd through 4th clonal stage data and raw mean of 5th clonal stage (Nursery/Infield stage).

Predictor	Cane yield	Stalk weight	Sugar content	Stalk number
APCH	0.24	0.53	0.61	0.75
PACH	0.24	0.53	0.59	0.76
AD	0.22	0.50	0.63	0.83
PS	0.26	0.55	0.63	0.86
LSM	0.28	0.52	0.66	0.82
BGa	0.30	0.54	0.72	0.74
BGb	0.30	0.54	0.73	0.71
BGTa	0.36	0.59	0.74	0.76
BGTb	0.35	0.58	0.74	0.75
BGTIma	0.35	0.57	0.70	0.74
BGTImb	0.42	0.63	0.73	0.74
Raw mean	0.08	0.48	0.18	0.59

Piepho (1998) comments about the advantage of computational simplicity of the analysis of means compared to a full model that incorporates replications. In this data set, it was not computationally feasible to incorporate genotype-by-environment (GE) effects using replicated data. Using genotype-trial means enabled the incorporation of a GE term into a predictive model dealing with numerous genotypes.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The resulting predictors, BGTIma and BGTImb, essentially equaled the best predictor for cane yield ($RMSE_{BGTIma}$ = 15.0 vs. $RMSE_{BGTa}$ = 14.9 Mg ha^{-1}; Table 3.4), but were not as low as the best predictors in the other traits. Weighting the predictor for the trial residual variance improved this predictor for all traits and, as earlier observed, appeared to improve the correlation to a small degree.

Using eight years of alfalfa (*Medicago sativa* L.) variety trials, Hill and Rosenberger (1984) compared check-based methods for combining germplasm data against three versions of BLUPs. They used a fixed 2-way analysis with trials and genotypes as factors and a cross-validation procedure to predict the performance of genotypes in a "left out" trial from the entire data set. The smallest average prediction error was obtained with the trial-heritability version of BLUP, which is equivalent to Henderson's (1977) procedure with variance components estimated from the unbalanced two-way analysis. No relationships were assumed among the genotypes (the different entries in the series of trials were considered unrelated in this study). If the set of genotypes being evaluated within a database are related, further advantages may be gained from BLUPs by using these genetic relationships in all the models with genotype regarded as a random factor (Panter and Allen, 1995). Genetic relationships among genotypes evaluated for yield in a few trials improved prediction in soybean (Panter and Allen, 1995). A genetic relationship-adjusted BLUP for genotype effects is worth considering at the early stages of breeding. This is because the large number of genotypes evaluated in these stages allows a good estimation of variance components. Poor variance component estimates, i.e., large standard errors of variance component estimates, are expected if the number of genotypes is very small. Furthermore, in early

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

selection stages there are few direct observations but likely a number of observations from sibling clones is large. Genotype by environment interaction was not modeled in the alfalfa and the soybean data-set analyses (Panter and Allen, 1995). Attempts to combine trials into a single estimate of yield for sugarcane would introduce interactions that are known to exist (Bull et al., 1992; Mirzawan et al., 1993).

This study used only plantcane data. In practice, sugarcane breeders use more data than used in this study to make their selection decisions. Although advancement decisions are made prior to harvesting the trials planted in the previous year, stalk numbers of the previous year's trials is recorded before the selection decisions have been made. Furthermore, trials are harvested three times, once each year. Although yield data are not independent among crops for a given plot, additional yield data are, however, provided by the ratoon crop (Milligan et al., 1996). A mixed model approach for combining data might account for the serial correlation among crop values.

Broad (overall) and narrow (environment-specific inference) inferences have been discussed in a mixed model context (MacLean, 1991). Even when at early stages of a breeding program, breeders are not feigning to predict performance of genotypes in specific environments. The incorporation of a GT term in the model is important because it modifies genotype BLUPs by better estimating the genetic and residual variances. Although the number of replications used in the selection stages involved in the analysis is not very different (two or three), the genetic and residual variances by stage demonstrated substantial range (Table 3.6), which is not surprising. The population is dramatically reduced by selection as it progresses from the pre-Nursery stages (1000 to 250 clones) to the Infield stage (20 to 30 clones).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3.6. Broad-sense genetic and residual variance components for three stages of selection in the Louisiana Sugarcane Variety Development Program.

Variance component	Cane yield	Stalk weight	Sucrose content	Stalk number
	$(Mg\ ha^{-1})^2$	$(g \times 10^{-2})^2$	$(g\ kg^{-1})^2$	$(no.\ m^{-2})^2$
		Infield stage		
σ_g^2	29.3	1.09	35.6	0.861
σ_e^2	85.8	0.91	65.7	0.932
		Nursery stage		
σ_g^2	67.3	1.73	47.7	0.792
σ_e^2	233.0	1.38	59.4	1.188
		pre-Nursery stage		
σ_g^2	120.3	3.51	75.1	0.808
σ_e^2	359.6	2.97	132.1	0.517

The methods of measuring yield and the range of environments also vary quite a bit. In the Infield stage, an entire plot is weighed compared to the earlier stages in which yields are estimated from stalk counts and sample stalk weights. The pre-Nursery stages are evaluated on two soil types but basically at only one location. The Nursery stage is planted at three locations, whereas the Infield stage is planted at one or two locations. The main testing location for the Infield stage is where the population was selected in the previous four stages. Hence one sees error variances for cane yield for instance, range from 85.8 $(Mg\ ha^{-1})^2$ in the Infield stages to 359.6 $(Mg\ ha^{-1})^2$ in the pre-Nursery stages. Such range suggests that a two stage-analysis to account for variance heterogeneity by appropriate weighting based on the reciprocals of the standard errors might be useful. Freshman (1997) successfully applied this estimation strategy in a wheat breeding program.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The analysis can be done by setting up a weight matrix **W** with the diagonal elements, r_i/s^2_i obtained from a previous ANOVA for each trial, and by replacing $\mathbf{R}=\sigma^2_e\mathbf{I}$ by $\mathbf{W}^{-1/2}\mathbf{R}\mathbf{W}^{-1/2}$ to generate a new variance-covariance matrix for the residual term in the model of phenotypic means.

The small differences between version *a* and *b* of BLUPs procedures [6] through [9] (Table 3.4.) suggested that BLUP accuracy was not dependent on check values. BLUPs for genotypes can be obtained also when check varieties fail or dramatically vary in component values under a mixed model approach. Multiple check varieties are used to provide backup in case of failure and to establish commercial comparison for minimum or maximum commercial productivity levels; e.g., CP74-383 was a high cane yield, low sucrose content-type check. On occasion, new commercial varieties may set dramatically higher standards. This was the case with the cultivar LCP85-384 (Milligan et al., 1994), which yielded 24 to 38% higher than the commercial varieties it replaced. Experimental cultivars tested against LCP85-384 and analyzed with the percent-of-the-check method [1] could not be realistically compared to experimental genotypes compared against older checks. This method (APCH, [1]) is the standard method used in the program. BLUP based methods provide a practical and usually better alternative to check based-predictors at least partly because of their freedom from this constraint.

4. CONCLUSIONS

Performance testing of new material is an important and expensive facet of all plant breeding operations. This study investigated the check-based predictor of *per se* (genotype) performance commonly used in the LSVDP. The simultaneous evaluation,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

obtained from the empirical validation based on LSVDP data, indicated that statistical methods exist that can improve prediction accuracy compared to the standard percent-of-the-check method [1] without increasing resource demands. BLUP procedures and standardization within a trial with respect to checks produced better predictions than those obtained by the average of percentage of checks. There are several other mixed models that could be investigated to maximize the accuracy of the estimate of the performance of a genotype from fewer evaluations. Research should be conducted to propose better models that are feasible from a computational point of view. All the BLUPs evaluated in this paper can be obtained using Proc Mixed/SAS (SAS Inst., 1997) run on PCs.

5. REFERENCES

Breaux, R. D., Fanguy H.P, Matherne R.J., Dunkelman P.H.. 1974. Registration of 'CP 65-357' Sugarcane. Crop Sci. 14: 605.

Bull, J.K., Hogarth D.M., Basford K.E. 1992. Impact of genotype x environment interaction on response to selection in sugarcane. Aust. J. Exp. Agric. 32,731-737.

Cochran, W.G. 1954. The combination of estimates from different experiments. Biometrics 10: 101-129.

Cullis, B.R., Thompson, F.M., Fisher, J.A., Gilmour, A.R., Thompson, R. 1996a. The analysis of the NSW wheat variety database. II. Modeling trial error variance. Theor Appl Genet 92:21-27.

Cullis, B.R., Thompson, F.M., Fisher, J.A., Gilmour, A.R., Thompson, R. 1996. The analysis of the NSW wheat variety database. II. Variance component estimation. Theor Appl Genet 92:28-39.

Fanguy, H. P., Breaux, R.D. 1981. Registration of 'CP 72-370' sugarcane. Crop Sci. 21: 798.

Fanguy, H. P., Garrison, D., Breaux, R.D. 1983. Registration of 'CP 74-383' sugarcane. Crop Sci. 23: 1220.

Fanguy, H. P., Dunkelman, P.H., Breaux R.D. 1979. Registration of 'CP 70-321' sugarcane. Crop Sci. 19: 413.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Frensham, A., Cullis, B., Verbyla, A. 1997. Genotype by environment variance heterogeneity in a two-stage analysis. Biometrics 53:1373-1383.

Harville, D.A.1990. BLUP Best linear unbiased prediction and beyond. *In* Gianola D., K. Hammond (eds.). Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, 239-276

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

Henderson, C.R. 1977. Prediction of future records. In E.Pollak et al. (ed.) Proc. Int. Conf. on Quantitative Genetics. Iowa State Univ. Press, Ames.

Hill Jr., R.R., Rosenberg, J.L. 1985. Models for combining data from germplasm evaluation trials. Crop Sci 25:467-470

Martin, F. A., Bischoff, K. P., Milligan, S. B., Quebedeaux, J. P., Dufrene, E. O., Hoy, J. W., Reagan, T. E, Giamalva, M. J., Miller, J. D., Breaux, R. D., Legendre, B. L. 1992. Registration of 'LCP 82-89' sugarcane. Crop Sci. 32: 499.

McIntosh, M.S. 1983. Analysis of combined experiments. Agron. J. 75:153-155.

McLean, R.A., Sanders, W.L., Stroup, W.W. 1991. A unified approach to mixed linear models. American Statistician 45:54-64.

Milligan, S. B. 1994. Test site allocation within and among selection stages of a sugarcane breeding program. Crop Sci. 34: 1184-1190.

Milligan, S. B., Martin, F. A., Bischoff, K. P., Quebedeaux, J. P., Dufrene E. O., Quebedeaux, K. L., Hoy, J. W., Reagan, T. E., Legendre, B. L., Miller, J. D.. 1994. Registration of 'LCP 85-384' sugarcane. Crop Sci. 34: 819-820.

Milligan, S. B., Gravois, K. A., Martin, F. A.. 1996. Inheritance of ratooning ability and the relationship of younger crop traits to older crop traits. Crop Sci. 36: 45-50.

Milliken, G.A., Johnson, D.E. 1989. Analysis of messy data. Volume 2: Nonreplicated experiments. New York: Van Nostrand-Reinold

Mirzawan P.D., Cooper M., Hogarth, D.M. 1993. The impact of genotype × environment interactions for sugar on the use of indirect selection in Southern Queensland. Aust. J. Exp. Agric. 33:629-638.

Panter, D.M., Allen, F.L. 1995. Using best linear unbiased predictions to enhance breeding for yield in soybean: II. Selection of superior crosses from a limited number of yield trials. Crop Sci. 35:405-410.

Piepho, H.P. 1998a. Stability analysis using the SAS System. Agron. J.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

SAS Institute. 1997. SAS/STAT software: changes and enhancements through release 6.12. SAS Inst., Cary, NC.

Searle, S.R. 1987. Linear models for unbalanced data. Wiley, New York

Searle, S.R., Casella,G., McCulloch, C.H. 1992. Variance components. Wiley, New York

Stroup, W.W. 1989. Why mixed models?. *In* Applications of Mixed Models in Agriculture and Related Disciplines. Southern Coop. Series Bull. No. 343. Louisiana Agric. Exp. Stn., Baton Rouge, Louisiana, 104-112.

White, T.L., Hodge, G.R., Delorenzo, M.A. 1986. Best linear prediction of breeding values in forest tree improvement. *In* Workshop of the Genetics and Breeding of Southern Forest Trees, Southern Region Information Exchange Group 40, Gainesville, Fl., 99-122.

Wolfinger, R.D., Federer, W.T., Cordero-Brana, O. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. Agron. J. 89:856-859.

Yates F, Cochran W.G.1938. The analysis of groups of experiments. J.Agric Sci, Cambridge, 28:556-580.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 4

INTEGRATING GENOTYPE-ENVIRONMENT COVARIANCE INTO THE COMPARISON OF GENOTYPE MEANS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1. INTRODUCTION

Replicated yield trials involving several environments are often used in late stages of breeding programs to select genotypes based on yield and other economically important traits. Each genotype is commonly tested in more than one environment represented by locations or years or their combinations. A usual feature of all multi-environment trials (MET) is the attempt to represent a relatively large target population of environments by a number of representative elements (Littell et al., 1996). In multi-environment trials, environments might be reasonably assumed as random effects (Piepho 1994). However, the genotype effects might be treated as fixed since only a few highly selected genotypes are usually involved in late breeding stages. Comparing genotype performance of new cultivars is the main aim of multi-environment yield trials in plant breeding. Two types of inference about genotype performance are of interest (1) broad inference – the general performance of a genotype, and (2) environment-specific or narrow inference – the performance of a genotype within a specific environment (McLean, 1991).

The traditional analytical approach for broad inference is based on genotype means that are subjected to multiple pairwise comparisons. Narrow inference from multi-environment trials relies on comparisons of genotypic means in specific environments (Littell et al., 1996). Unfortunately, this procedure does not use all the available information. It is only possible to infer about performance in a specific environment for genotypes that have been tested in that environment. Mixed model prediction may use information from an entire data set to obtain environment-specific inferences.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The need to identify genotypes specifically adapted to some target environments (environment-specific genotype recommendations) has prompted extensive research about genotype-by-environment interaction (GE) (Kang, 1990; Kang and Gauch, 1996). The stability approach to address GE (Lin et al., 1986; Becker and Leon, 1988; Crossa, 1990; Lin and Binns, 1994; Kang and Gauch, 1996), which has been used for simultaneous selection for yield and stability, has been regarded as beneficial for breeders, official test stations, and growers (Weber et al., 1996). Most of the analytical procedures to quantify a genotype's contribution to the overall GE are based on a fixed effects model approach. Such fixed models are applicable only to balanced data. Kang and Magari (1996) used the restricted maximum likelihood (REML) method under a mixed model to estimate stability variances in unbalanced data sets when analyzing GE for ear moisture loss rate in corn (*Zea mays* L.). The REML variance components, assignable to each genotype, estimate the same parameters as Shukla's stability variance (Shukla, 1972). The mixed model with heterogeneous (by genotype) GE terms is *a priori* more tenable than the traditional mixed analysis of variance in the sense that it allows different stability parameters for each genotype but it assumes independence among the GE effects.

By further modeling the variance-covariance structure of environment and interaction random effects, well known stability measures can be expressed as parameters of closely related mixed models (Piepho, 1998a). The common regression approach for studying genotype sensitivities to environmental changes with multiplicative models for the GE terms (Yates and Cochran, 1938; Finlay and Wilkinson 1963; Eberhart and Russell, 1966) can be handle by integrating a factor

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

analytic variance-covariance structure into a mixed model for the observed yield (Oman, 1991; Piepho, 1997; Piepho, 1998b).

Among the analytical methods involving multiplicative interaction, the "Additive Main effects and Multiplicative Interaction" (AMMI) models have been widely used because the GE can be interpreted in more than one dimension (Vargas, 1998). In AMMI models, the interaction terms are explained by the sum of multiplicative functions of genotype and environment scores (Gauch, 1988; Zobel et al., 1988). The AMMI models for analyzing GE were proposed in relation to former ideas of modeling interaction in factorial experiments (Williams, 1952) where the factors, in this case environments and genotypes, are assumed to be fixed. In the fixed model framework, the genotype and environment score vectors (principal components, PCs) are obtained from the singular value decomposition (SVD) of the matrix containing the residuals after adjusting the data for environment and genotype main effects (Mandel, 1971). The resulting genotype and environment scores are commonly visualized in biplots (Gabriel, 1971; Price and Shafii, 1993). Biplots from AMMI models are useful tools in plant breeding because they allow the identification of genotypes that show smaller interaction with environment and higher yield values. They can also identify genotypes that perform well at specific sites (Yau, 1995; Shafii and Price, 1998). SVD strictly requires a complete data set (observation of all genotypes within all environments). However, a common feature of yield trials is that lists of entries vary from year to year because new entries are included as they become available and those with poor performance are deleted from further consideration (Hill and Rosenberg, 1985). The deletion and substitution results in unbalanced data. Even within a year, it is rare to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

have balanced data since some replications or locations may not be planted with all genotypes.

Mixed model and restricted maximum likelihood-based estimation procedures for the parameters in the models (Searle et al., 1992) provide a more flexible analytical approach for the analysis of multi-environment trials because balanced data are not required (Hill and Rosenberg, 1985; Stroup and Mulitze, 1991; Piepho, 1994, 1997, 1998a). Mixed model analysis basically models the underlying covariance structure.

In particular, the regression approach and AMMI mixed model analysis are based on a covariance matrix for the genotypic means within an environment, with features of the factor analytic type of variance-covariance structure (Jenrich and Schluchter, 1986; Denis et al., 1996; Piepho, 1997). They account for possible correlations among the interaction terms, which can be realistically expected. When some environments or some genotypes are correlated, the GE terms involving those environments and genotypes may be correlated.

Nowadays, it is possible to apply mixed models with factor analytic and even more sophisticated, variance-covariance structures in SAS (SAS Inst., 1997) and other statistical packages with mixed model applications. However, the biplots are not readily obtained from regular outputs. In addition, the interpretation goals are the same whether one use an AMMI mixed model or a traditional AMMI (fixed) model; the parameter-types used to identify interaction patterns change, however.

The purpose of this study was to compare five classes of mixed models for analyzing multi-environment yield trials under a unified approach. The development of mixed model AMMI derived biplots was a related goal.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. MATERIALS AND METHODS

Mixed models involving parameters analogous to Shukla's stability variances, Eberhart and Russell's sensitivity coefficients, and genotype-environment scores, as in the fixed AMMI models, were compared with the simplest mixed model that assumes homogeneous and independent GE terms with and without homogeneous variance for the error terms. The traditional fixed model approach was included for reference. Parameter estimates and a proposed procedure to obtain biplots under a mixed AMMI model was illustrated with a set of sugarcane (*Saccharum* spp.) multi-environment yield trials .

2.1. Models for Multi-Environment Trials

The models employed in this study to analyze yield trials use a variation of the following model,

$$y_{ijk} = \mu + E_i + R_{k(i)} + G_j + GE_{ji} + \varepsilon_{ijk}$$

where y_{ijk} is the *k*-th observation for the *j*-th genotype in the *i*-th environment, μ , E_i, G_j, $R_{k(i)}$, GE_{ji} denote the overall mean, the environmental effect [i =1,...,s], the genotype effect [j =1,...,g], the replication-within-environment effect [k =1,...,r], and the genotype-by-environment interaction effect, respectively; ε_{ijk} is the error term associated with y_{ijk}. All models assumed genotype effects as fixed. In addition to the regular fixed model which considers all effects, except the error term, as fixed the models used in this study assumed environments, blocks-within-environment effects, interaction and the error terms as random. The assumptions for the random effects are: environmental

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

effects, E_i, are iid $N(0, \sigma^2_E)$ and replication effects, $R_{k(i)}$, are iid $N(0, \sigma^2_R)$. The GE terms are also regarded as normal random effects with zero means but with a variance-covariance matrix not necessarily implying independence and homogeneity of variances. Error terms are assumed to be iid $N(0, \sigma^2_e)$ for all models, except model [5] where the residual variance is allowed to be different at each environment. Environmental, replication, interaction and error effects are independent of one another.

As a direct consequence of the model assumptions, the variance of the yield values is the sum of the variances of each random effect. For simplicity, I assumed independence not only of error term but also of environmental and replications-within-environment effects. The assumption implies that the environments provide independent information. Although one is assuming that environment effects are not correlated, the response means within a given environment will be correlated because of the type of variance-covariance matrix associated with the mixed model. The means of any two genotypes in a specific environment, y_{ij} and $y_{ij'}$, have the covariance

$$\text{Cov}\,(y_{ij}, y_{ij'}) = \sigma^2_E + \text{Cov}(GE_{ij}, GE_{ij'}), \text{ for } j \neq j'$$

The mixed models evaluated in this study varied in the variance-covariance structure imposed on the interaction term, $\text{Cov}(GE_{ij}, GE_{ij'})$ (Table 4.1).

Model [1] (MIXED ANOVA) assumed that the GE terms have the same variance and are independent. Model [2] (MIXED SHUKLA) enabled GE terms to have different variances but assumes they are independent. The model [2] assumes that all GE terms involving a particular genotype have the same GE variance, thus there will be

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

as many different GE variance components as the number of genotypes, so they are analogous to Shukla's stability variances. Hence the name given to the model even though Shukla (Shukla, 1972) did not express stability variances as parameters of a mixed model.

Model [3] (MIXED AMMI) considers multiplicative GE effects,

$$GE_{ij} = \sum_{m=1}^{M} \lambda_{mj} x_{mi} + d_{ij}$$

where the first part ($\sum_{m=1}^{M} \lambda_{mj} x_{mi}$) is the sum of multiplicative terms used to explain interaction signals and d_{ij} is the residual interaction term. Each multiplicative term represents a linear regression model of the residuals from the main effect model for the *j*-th genotype on a latent unobservable variable related to the *i*-th environment. A sum of multiplicative terms is used to model GE variability pattern in more than one dimension. The subscript *m* indexes the axis of variability on which the fixed genotype and random environment scores are obtained. Thus, for each axis of variation, the genotypic score $\lambda_{j.}$ can be interpreted as the response of the *j*-th genotype to changes in some latent environmental variable with value x_i in the *i*-th environment. The model for the GE terms resembles the non-additive part of the traditional AMMI models (Gauch, 1988; Zobel et al., 1988), but in the fixed AMMI models, environment scores are fixed. The sum of multiplicative terms is part of the expected value of y_{ijk} in the fixed approach, whereas under the mixed model, this belongs to its covariance structure. The models imposed on the GE terms lead to specific variance-covariance matrix types for the vector $\mathbf{y}_{(i)}$ containing the genotypic means in the *i*-th environment (Table 4.1).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Model [4] (MIXED E&R) does not contain the main effect for environment and also considers multiplicative GE effects,

$$GE_{ij} = \lambda_j x_i + d_{ij}$$

where λ_j is the sensitivity of the *j*-th genotype to a non-observed environmental variable x_i and d_{ij} is the unexplained part of the genotype-by-environment interaction. The deviations d_{ij} are allowed to have a separate variance for each genotype, $\sigma^2_{d(j)}$. Despite the fact that environmental variable is assumed as random, the model resembles the Eberhart-Russell (1966) regression model to study genotype-by environment interaction. A genotype with large λ_j absolute value of shows a large sensitivity to changes of the underlying random variable x_i .

Model [5] (MIXED HetR) is the same that model [1], i.e. assumed that the GE terms have the same variance and are independent, but allows for heterogeneous by environment residual variances, $R= \sigma^2_{e(j)}I$.Model parameters were estimated by REML (Searle et al., 1992). All calculations including fixed and random effects estimates were done using Proc Mixed/SAS (SAS Inst., 1997) which solved the mixed model equations on the REML estimates. The decision about the appropriate number of multiplicative terms to use in the MIXED AMMI was based on the difference between –2 Residual Log Likelihood (-2 Res LL) of nested AMMI models, i.e., AMMI with 1,2,3... multiplicative terms. For example, D= (-2ResLL(AMMI(1))) – (-2ResLL(AMMI(2))) was employed as a statistic to evaluate the need of incorporating a second multiplicative term in an AMMI model with one multiplicative component.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.1. Mixed models employed in the analysis of multi-environment yield trials.

Model	Model Equation†	Interaction effects Assumptions	Covariance structure for $\mathbf{y}_{(i)}$‡
[1] MIXED ANOVA	$y_{ijk} = \mu + E_i + R_{k(i)} + G_j + GE_{ij} + \varepsilon_{ijk}$	$GE_{ij} \sim$ iid $N(0, \sigma^2_{GE})$	Σ**env** $= \mathbf{J}\sigma^2_E + \mathbf{I}\,\sigma^2_{GE}$
[2] MIXED SHUKLA	$y_{ijk} = \mu + E_i + R_{k(i)} + G_j + GE_{ij} + \varepsilon_{ijk}$	$GE_{ij} \sim$ iid $N(0, \sigma^2_{GE(j)})$	Σ**env** $= \mathbf{J}\sigma^2_E + \mathbf{I}\,\sigma^2_{GE(j)}$
[3] MIXED AMMI	$y_{ijk} = \mu + E_i + R_{k(i)} + G_j + GE_{ij} + \varepsilon_{ijk}$	$GE_i \sim N(0, \sum_{m=1}^{M} \lambda^2_{mj} + \sigma^2_d)$ for all i.	Σ**env** $= \mathbf{J}\sigma^2_E + \Lambda\Lambda' + \mathbf{I}\,\sigma^2_d$
	$GE_{ij} = \sum_{m=1}^{M} \lambda_{mj} x_{mi} + d_{ij}$	$Cov(GE_{ij}, GE_{ij'}) = \sum_{m=1}^{M} \lambda_{mj}\lambda_{mj'}$ for $j \neq j'$	
[4] MIXED E&R	$y_{ijk} = \mu + R_{k(i)} + G_j + GE_{ij} + \varepsilon_{ijk}$ $GE_{ij} = \lambda_j x_i + d_{ij}$	$GE_i \sim N(0, \lambda^2_j + \sigma^2_{d(j)})$ for all i. $Cov(GE_{ij}, GE_{ij'}) = \lambda_j\lambda_{j'}$ for $j \neq j'$	Σ**env** $= \Lambda\Lambda' +$ **diag**$(\sigma^2_{d(j)})$
[3] MIXED HetR	$y_{ijk} = \mu + E_i + R_{k(i)} + G_j + GE_{ij} + \varepsilon_{ijk}$	$GE_{ij} \sim$ iid $N(0, \sigma^2_{GE})$	Σ**env** $= \mathbf{J}\sigma^2_E + \mathbf{I}\,\sigma^2_{GE}$ $\mathbf{R} = \sigma^2_{e(i)}\mathbf{I}$

† μ: overall mean; E_i: random environment *i* effect; $R_{k(i)}$: random replication-within-environment effect; G_j: fixed genotype *j* effect, GE_{ij} random genotype-by-environment interaction; λ_{mj} ($j = 1,\ldots,g$) genotype factor loading on the *m*-th multiplicative interaction term, x_{mi} : *m*-th predicted score for a latent environmental variable in environment *i*; d_{ij} : residual interaction term; ε_{ijk} error terms associated to the response y_{ijk}. R is the variance-covariance of the error terms.

‡ $\mathbf{y}_{(i)}$: vector of genotype means in environment *i* ; Σ**env** : variance-covariance matrix of $\mathbf{y}_{(i)}$; **J**: *gxg* matrix of 1's; **I**: *gxg* identity matrix; Λ: *gxM* matrix of genotype factor loadings for each multiplicative term $m = 1,..,M$.

The likelihood ratio-based test (LRT) obtained by comparing those likelihoods was also used to compare models [2] to [5] against the simplest mixed model [1]. Differences, D, between -2ResLL were compared with a χ^2 variable with degrees of freedom equal to the difference in the number of covariance parameters between the two models being compared.

Generalized least squares means for each genotype were used for broad inference. Pairwise comparisons among these genotype means used a sampling error variance for the mean difference that incorporates all covariance parameters (Littell et al., 1996). I used the SAS macro 'pdmixmac612' (SAS Inst., 1997) to align the means, obtained in Proc Mixed/SAS, in accord with the significance of pairwise multiple comparisons (Appendix B). BLUPs (Searle et. al., 1992) were used to predict the performance of genotype *i* in environment *j* (narrow inference). For narrow inference under a mixed model, one is interested in the BLUP of the conditional expectation μ_{ij}, BLUP(μ_{ij}). BLUP(μ_{ij}) is a linear combination of the estimated genotype mean for genotype *j* (estimated fixed effect) and the estimated random effects for environment *i* and the GE term *ji*, BLUP(E_i) and BLUP(GE_{ji}), respectively .

Mixed AMMI models were employed to analyze GE_{ji} by constructing biplot representations. A MIXED AMMI is essentially a multi-level factor analysis (Gollob, 1968) model with M levels. Latent factors at level *m* (m=1,...,M) represent environmental random variables. I deduced the values of those random environmental scores, x_{mi}, by pre-multiplying the vector of BLUPs of GE terms in environment *i*,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

obtained in Proc Mixed/SAS, to the inverse of the estimated loading matrix Λ that can be constructed also from the regular Proc Mixed/SAS output.

Biplots representing GE variability in two dimensions, i.e., two multiplicative terms, were constructed by superimposing the standardized genotype factor loading on multiplicative terms m=1 and m=2 with random environmental scores on the same multiplicative terms. To facilitate simultaneous interpretation of yield values and GE, genotype scores on the first and second multiplicative terms are plotted against trait mean values. Appendix D contains a program to obtain biplots for a MIXED AMMI.

2.2. Data and Validation Procedure

In the Louisiana Sugarcane Variety Development Program (LSVDP) replicated tests culminate in outfield trials (Milligan, 1994). Outfield trials usually overlap experimental material from different series with check or commercial varieties. For example, varieties one to eight (Table 4.2) were check commercial varieties in the outfield trials conducted between 1996 and 1998. Regular outfield tests involves 10 to 12 genotypes per trial. Trials are conducted at several (7 to 10) commercial farms distributed throughout the 158 000-ha crop region. Each trial is laid out in a randomized complete-block design with three replications and use 53.5 m^2 (three 1.8m wide rows by 9.7m long) plots. To compare prediction accuracy of different mixed models, I used both components of sugar yield, i.e., cane yield (Mg ha^{-1}) and sucrose content (g sucrose kg^{-1}cane). The data set used Louisiana advanced variety trial plantcane data (outfield tests) from 1996 to 1998 (Quebedeaux et al., 1996, Quebedeaux et al., 1997, Guillot et al., 1998).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.2. Sugarcane yield trials conducted in Louisiana across five years (1996-1998). Codes for participating genotypes (varieties) and environments (farms).

Test year	Variety Code†	Farm Code‡
1996	1,2,3,5,6,7,8,9,10,11,12,16	1,2,4,5,6,7,8,9,10
1997	2,3,4,5,6,7,8,10,11,14,15	2,3,4,5,6,8,10
1998	2,7,8,10,14,17,18,19,20,21	1,2,3,5,6,10,11

† 1:All.-A.V.Allain & Sons, 2:B.S.-Bon Secour, 3:Geo.-Georgia, 4:Gln.-Glenwood, 5: Lan.-Lanaux, 6: Mag.-Magnolia, 7:Oak.-Oaklawnhy, 8: P.A.-Palo Alto, 9:R.L.-Raceland, 10: R.H.-Ronal Hebert, 11:StJ.-Levert-St. John.
‡ 1:CP65-357, 2:CP70-321, 3:CP72-370, 4:CP79-318, 5:LCP82-089, 6:LHo-LHo-83153, 7:LCP85-384, 8:HoCP85-845, 9:HoCP91-552, 10: HoCP91-555, 11:LHo92-314, 12:L92-315, 13: HoCP92-618, 14: HoCP92-624, 15: HoCP92-648, 26: HoCP92-674, 17: HoCP93-754, 18:L94-426, 19:L94-428, 20:L94-432, 21:HoCP94-806.

Predictive accuracy of narrow inference from models (Table 4.1) was obtained by a "leave-one-block-out" cross-validation procedure (Appendix C). Independent cross-validation was run for each test year. For each outfield trial, the data set was split into two subsets, one with two replications per environment (calibration data) and the other with one replication per environment (validation data). The calibration data set was used to predict variety performance in each environment. Predicted performance was compared to the observed yield for each variety in the validation data set. The process was repeated 30 times for different randomizations, i.e., different sets of two blocks per environment. The average of squared differences between predicted and observed values of each genotype were used to approximate prediction accuracy of narrow inference. At each iteration of the validation procedure, counts were made for each environment to ascertain how many of the top 50% varieties in the validation set would also have been ranked in the top 50% of a variety list sorted in accord with the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

variety BLUPs for that particular environment. The average percent of varieties in the top 50% of both lists is denoted by P (50|50).

3. RESULTS AND DISCUSSION

Cane yield data revealed significant GE variance components for each year in the sugarcane yield trials from 1996 to 1998 (Table 4.3). Significant GE for sucrose content (α=0.05) was observed only in 1997.

Table 4.3. Variance component estimates for environmental (E) and genotype-by-environment (GE) effects for three years (1996 to 1998) of sugarcane variety trials

Variance Component	Cane Yield $(Mg\ ha^{-1})^2$	Sucrose Content $(g\ sucrose\ kg^{-1}cane)^2$
	----------1998----------	
σ^2_E	21.68	24.73
	(0.166) †	(0.113)
σ^2_{GE}	25.58	8.83
	(0.001)	(0.057)
σ^2_e	41.319	40.38
	----------1997----------	
σ^2_E	26.02	134.98
	(0.116)	(0.088)
σ^2_{GE}	10.65	10.62
	(0.012)	(0.010)
σ^2_e	33.191	31.88
	----------1996----------	
σ^2_E	30.80	81.52
	(0.060)	(0.055)
σ^2_{GE}	11.67	1.12
	(0.001)	(0.667)
σ^2_e	31.77	39.02

† In parenthesis P-values for the hypothesis of variance component equal to zero (Z test).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The variance component standard errors obtained from REML procedures and Z tests for the hypothesis that the variance components equal zero are only asymptotically valid. Poor approximations should be expected when dealing with variance components estimated with a small degrees of freedom, such as with σ^2_E. Tests for a common σ^2_{GE} are more reliable. The F-test P-value of fixed variety effects was smaller than 0.001 for both traits in each multi-environment trial.

Least square means (LSMeans) are commonly used for broad genotypic inferences across environments. Standard errors for each LSMeans were computed using the general formula for the variance of an estimable function under the mixed model. Therefore, the variance of an estimated genotype mean involves the variances of all random effects in the model (Littell et al., 1996). Because of the differences in variance-covariance structures, genotype groupings differ among the approaches relative to different models. The larger the GE , the larger the expected grouping differences. The GE in this sugarcane data set is not as important as is often observed in variety trials of other crops (Kang and Gauch, 1996) or previously reported GE interactions for sugarcane (Kang and Miller, 1984). However, different assessments of the yield performance of varieties across a range of environments were observed in each MET. The genotype mean separation obtained using 1998 MET data and different models, is presented in Tables 4.4 and 4.5 for cane yield and sucrose content, respectively. Differences in standard errors among genotypes means for mixed models [2] to [4] are due to difference in GE response. The mixed model approach to stability

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

analysis allows one to obtain genotype (broad) mean separations that combine yield measures and stability parameters.

If the data are completely balanced and the residual variance is assumed to be constant, then model [1] should assign the same variance to all observations. The data set of this study contained three replications per variety at most of the farms, but for some variety-environment combinations there were only two replications. Slight differences in standard errors for the broad inference means under model [1] reflect this unbalance.

Models [2] to [4], as expected, showed larger differences among the standard errors of the genotypic means. For example, for 1998 data both Mixed Shukla [2], Mixed AMMI model [3], and Mixed E&R [4] assigned larger standard errors to varieties L94-428, LCP85-384, and CP70-321 than to other varieties. The stability variance estimates (GEj) for L94-428, LCP85-384, and CP70-321 were 35.17, 63.88, and 88.74 $(Mg\ ha^{-1})^2$, respectively, whereas the average variance for the remaining varieties was about 13.00 $(Mg\ ha^{-1})^2$. Thus, a larger standard error was used for mean separation of genotypes involving larger GE interaction components. Because the number of environments is not large in this data set (seven farms in 1998), the stability parameters $\sigma^2_{GE(j)}$ for $j = 1,\ldots,10$ are estimated with large standard errors and hence the Z test for the covariance parameters was not employed. The variance components reported above should be interpreted only for a tentative ranking of the stability of those varieties in 1998. When the number of environments is large and greater than the number of genotypes, the Z test may be employed to test if those values reflect a genotype feature.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Multiplicative interaction models in addition to models that use a Shukla-type variance component in the GE also provide extended standard errors. The parameters related to genotype stability in multiplicative models (genotype factor loadings) such as those of Eberhart and Russell (1966) and AMMI models (Gauch, 1988) are covariance parameters, i.e, they make up the variance and covariance of the data. They can be used to visualize GE interaction but also to construct an extended standard error for genotype mean separation integrating genotype yield and stability. The product of the genotype loadings estimates the covariance between genotype GE effects of pairs of genotypes in a particular environment (Table 4.1). Comparison of genotype means under a multiplicative model like models [3] and [4] will account for this aspect of GE. A multiplicative model for GE implies that the difference of two genotypes in a particular environment depends on the differences of genotypes scores and the magnitude of the environmental random scores predicted for the particular environment.

The approach, treating all factors as fixed, compares genotype means without regard to GE variances. It only accounts for residual error variances. The reported standard errors for genotype mean comparisons were smaller under the fixed approach than the mixed models because the fixed model ignores variability due to GE (Tables 4.4 and 4.5). The mixed models model [5] produced a finer separation than the mixed model [1] indicating the importance of controlling for heterogeneous residual variance in the LSVDP outfield trials when modeling cane yield data. Differences in mean separation among models were smaller for sucrose content, probably, because of the smaller GE interaction associated with this trait.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The relatively small GE observed in these trials is likely a direct result in the pre-outfield multi-location testing (Milligan, 1994). Successful varieties must display high yields across all tested environments to be advanced to the outfield trials. Hence, they have been screened for low GE prior to testing in the outfield trials.

Comparing likelihoods among the mixed models allows one to select the most appropriate model from among those under consideration. Likelihood ratio tests are obtained by comparing the differences between the quantities "-2 Residual Log Likelihood" for a given pair of models. The difference can be compared against a χ^2 with degrees of freedom equal to the difference in the number of GE covariance parameters between the models. For example using 1998 data, to compare the simplest Mixed ANOVA against the Mixed Shukla model, i.e model [1] vs. model [2], the difference between the respective functions of likelihoods, D=1453.21-1442.53 = 10.68, is compared to a with χ^2 variable from a distribution with 14-4=10 degrees of freedom. A lower residual likelihood score is better than a higher score. The results of comparing the four mixed models with homogeneous residual variance indicate that an AMMI model [3] with two multiplicative terms (AMMI(2)) was better model for cane yield data than was the simple homogeneous and independent GE model [1] for test years 1998 and 1997 (Table 4.6). However, improvements over the Mixed ANOVA [1] were non-important for 1996. In this data, the Mixed Shukla [2] and the Eberhart and Russell [3] model did not improve, over the Mixed ANOVA [1], the model for cane yield.

These results suggest that modeling the correlation between interaction terms may be a good strategy when analyzing cane yield trials at the LSVDP.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.4. Cane yield least squares means† using five mixed modeling approaches for1998 sugarcane yield trials.

Variety	Mixed ANOVA [1]	Mixed Shukla [2]	Mixed AMMI(2) [3]	Mixed E&R [4]	Mixed HetR [5]	Fixed Model
	Least Squares Means (Mg ha^{-1})					
L94-428	69.20 A ±3.00	68.99 AB ±4.18	68.88 AB ±4.32	68.99 AB ±4.27	70.37 A ±2.77	68.91 A ±1.45
LCP85-384	68.56 A ±2.97	68.57 AB ±3.72	68.51 AC ±3.88	68.57 AB ±3.92	69.61 AB ±2.73	68.66 A ±1.41
HoCP91-5555	68.29 AB ±2.98	68.28 A ±2.84	68.29 AC ±2.76	68.28 A ±3.09	68.19 ABC ±2.75	68.29 A ±1.40
HoCP94-806	66.72 AB ±2.98	66.72 A ±2.17	66.72 AB ±2.32	66.72 A ±2.50	66.10 ABCD ±2.75	66.72 A ±1.40
L94-426	65.88 AB ±2.98	65.88 AB ±2.53	65.90 ABC ±2.47	65.88 AB ±2.55	65.07 ABCDE ±2.75	65.80 ABC ±1.41
HoCP92-624	65.77 ABC ±3.03	65.69 AB ±2.22	65.88 ABC ±2.75	65.70 AB ±1.77	65.31 ABCDE ±2.78	65.88 AB ±1.40
HoCP85-845	65.16 ABC ±2.95	65.09 ABC ±2.68	64.87 ABCD ±2.59	65.12 ABC ±2.85	64.34 BCDE ±2.71	65.41 ABC ±1.35
CP70-321	62.82 ABC ±2.95	62.83 ABC ±3.10	62.92 BDE ±3.21	62.83 ABC ±3.04	62.50 CDE ±2.71	62.87 BCD ±1.35
L94-432	61.80 BC ±3.00	61.84 BC ±2.33	61.78 CDE ±2.61	61.84 BC ±2.16	62.00 DE ±2.77	61.67 CD ±1.45
HoCP93-754	59.08 C ±2.98	59.08 C ±3.08	59.08 D ±3.07	59.08 C ±3.13	60.02 E ±2.75	59.08 D ±1.45

† Means followed by the same letter following are not significantly different from each other at the α=0.05.

‡ For model description see Table 4.1. 'Fixed Model' is two way factorial model with all effects as fixed.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.5. Sucrose content least squares means† using four mixed modeling approaches for1998 sugarcane yield trials.

Variety	Mixed ANOVA [1]	Mixed AMMI(2) [3]	Mixed E&R [4]	Mixed HetR [5]	Fixed Model
	Least Squares Means (g sucrose kg^{-1} cane)				
L94-432	139.92 ±2.61 A	139.92 ±2.21 AB	139.94 ±2.76 A	140.12 ±2.61 A	139.86 ±1.47 A
L94-428	138.93 ±2.59 AB	138.96 ±2.37 AC	138.93 ±2.25 AB	138.81 ±2.59 AB	138.93 ±1.42 AB
LCP85-384	136.22 ±2.57 ABC	136.03 ±3.53 AC	136.21 ±3.44 ABC	136.11 ±2.58 AB	136.35 ±1.42 ABC
HoCP94-806	135.98 ±2.59 ABC	135.98 ±2.05 AB	135.96 ±1.87 ABC	136.03 ±2.59 AB	135.98 ±1.42 ABC
CP70-321	135.83 ±2.55 ABC	135.69 ±2.35 ABC	135.69 ±2.35 ABC	135.63 ±2.56 AB	135.96 ±1.37 ABC
HoCP91-5555	135.45 ±2.59 ABC	135.45 ±2.03 ABC	135.45 ±1.73 ABC	135.53 ±2.59 AB	135.45 ±1.41 BC
L94-426	134.83 ±2.59 ABC	134.83 ±2.62 ABCD	134.83 ±2.37 C	134.76 ±2.59 B	134.83 ±1.42 C
HoCP93-754	134.76 ±2.59 BC	134.76 ±3.49 BDE	134.76 ±3.28 ABC	134.53 ±2.59 B	134.76 ±1.42 C
HoCP92-624	133.46 ±2.63 C	133.31 ±2.10 CDE	133.21 ±2.39 BC	133.84 ±2.63 B	133.33 ±1.53 C
HoCP85-845	125.94 ±2.57 D	125.91 ±2.71 D	126.07 ±3.19 D	126.03 ±2.56 C	125.99 ±1.37 D

† Means followed by the same letter are not significantly different from each other at the α=0.05.

‡ For model description see Table 4.1. REML algorithm did not converge for a Mixed Shukla model.

Correlation among locations might affect correlations among the GE terms. AMMI(3) models were not necessary to model interaction patterns in this data set (results not shown).

Comparisons of model fitting information for sucrose content indicated that an AMMI model should be preferred for analyzing 1998 and 1997 data, whereas in 1996, the simplest MIXED ANOVA might be adequate. AMMI models with two multiplicative terms were not suitable for sucrose content. This may be related to the relatively smaller GE interaction for sucrose content than for cane yield in the data sets.

Table 4.6. Model fitting information for five models employed to analyze cane yield and sucrose content in three years of multi-environment sugarcane trials

Model	Number of Covariance Parameters	-2 Residual Log Likelihood		
		1998	1997	1996
		for cane yield		
Mixed ANOVA	4	1453.21	1496.46	2039.06
Mixed SHUKLA	4+g	1442.53	1460.10	NA
Mixed AMMI(1)	4+g	1444.62	1460.97	2031.87
Mixed AMMI (2)	4+g+(g-1)	1414.53	1442.81	2010.70
Mixed E&R	3+2*g	1440.08	1460.02	2030.02
Best model		AMMI(2)	AMMI(2)	ANOVA
		for sucrose content		
Mixed ANOVA	4	1416.08	1468.17	2068.72
Mixed SHUKLA	4+g	NA	1450.45	2063.55
Mixed AMMI(1)	4+g	1396.98	1432.44	2055.21
Mixed AMMI (2)	4+g+(g-1)	1388.98	1416.09	2047.71
Mixed E&R	3+2*g	1397.35	1448.12	2061.52
Best model		AMMI(1)	AMMI(1)	ANOVA

† NA: non available because of REML algorithm does not converge

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Biplot analysis can often provide better insight into genotype by environment interaction responses than means or GE test alone. As an example, the 1998 cane yield and sucrose content GE scores were plotted. The biplot representing the genotype and environment scores on the two multiplicative terms for a Mixed AMMI(2) indicates that genotypes CP70-321 (2), LCP85-384 (7) and L94-428 (19) contributed more to the cane yield GE variability in 1998 than the other varieties (Fig 4.1). This is indicated by the fact that the genotype scores are far from the origin of either axis. The relative poor yield of genotype 19 in environment 6 contrasts the very high yields observed environments such as 2,10, and 11 (Table 4.7). Genotypes 2 and 7 yield performances were negatively correlated with environment 5 effects whereas genotype 19 was negatively correlated to environment 6. They are situated in opposite diagonal quadrants in the biplot.

Table 4.7. Cane yield from sugarcane variety trials conducted in 1998

Variety	Code	Environments						
		1	2	3	5	6	10	11
		---------- Mg ha^{-1} ----------						
CP70-321	2	58.29	62.85	65.01	51.49	76.83	67.55	58.05
LCP85-384	7	70.92	75.84	70.66	46.08	76.16	73.39	67.03
HoCP85-845	8	57.70	74.18	68.11	67.42	71.04	65.48	53.92
HoCP91-555	10	71.54	76.66	66.80	62.09	80.42	60.74	59.74
HoCP92-624	14	60.13	68.48	68.00	63.77	73.46	66.36	62.43
HoCP93-754	17	54.54	73.12	66.34	50.05	51.08	59.60	58.79
L94-426	18	57.14	78.24	65.19	70.42	65.08	63.93	61.12
L94-428	19	67.64	87.06	65.89	57.23	53.96	79.16	70.25
L94-432	20	61.42	73.83	57.94	59.53	58.68	60.27	62.24
HoCP94-806	21	63.60	77.54	61.05	62.16	73.94	63.02	65.70

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Plotting the genotype and environment scores for the first multiplicative interaction terms against cane yield values, it is possible to observe that variety 19 (experimental), with a high yield across environments, is one of the most unstable variety (Fig 4.2). Varieties 7 and 10 (both commercial varieties) followed in yield but with more stable performance.

Sucrose content GE variability was much smaller than that for cane yield GE. One multiplicative term adequately explained GE (Table 4.6). In 1998, varieties 17, 7, 14 and 10 displayed relatively high GE responses (Fig 4.3). Sucrose contents for sugarcane variety trials conducted in 1998 are shown in Table 4.8.

Although the estimation of genotype and environment scores under a fixed model involves procedures that are quite a bit different than those under a Mixed AMMI model, both approaches lead to similar interpretations from graphical representations. Biplots of the first principal GE component versus sucrose content (fixed AMMI model) (Fig. 4.4) suggested similar conclusions about variety yield potential and stability as indicated by the mixed AMMI model (Fig. 4.3). If factor loadings as estimated by SAS are multiplied by −1, the same representation is obtained for both fixed and mixed AMMI-based procedures. Although biplots derived from the fixed model approach, i.e., using the SVD could be obtained for 1998 since the MET in that year was balanced, they cannot be used to analyze other METs or in an analysis combining METs across years. Hence, using a mixed AMMI approach provides a decisive functional advantage to the fixed AMMI approach.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.8. Sucrose content for sugarcane variety trials conducted in 1998.

Variety	Code	Environments						
		1	2	3	5	6	10	11
		kg sucrose Mg^{-1} cane						
CP70-321	2	131	134	135	147	140	137	129
LCP85-384	7	122	139	143	150	138	134	130
HoCP85-845	8	117	122	133	136	136	117	120
HoCP91-555	10	135	126	140	137	137	136	137
HoCP92-624	14	137	128	142	135	134	133	123
HoCP93-754	17	119	135	140	147	136	131	136
L94-426	18	128	140	138	142	137	130	129
L94-428	19	138	143	139	148	141	129	135
L94-432	20	139	140	145	144	150	129	133
HoCP94-806	21	136	134	138	141	142	131	130

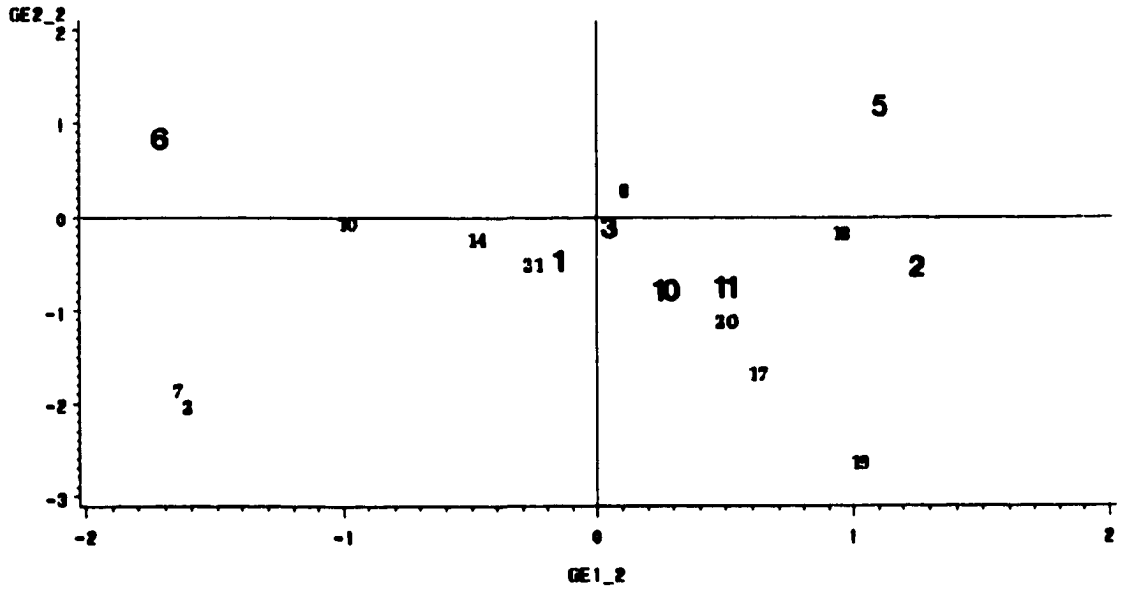

Figure 4.1. Genotype and environment scores on first and second multiplicative term of a Mixed AMMI (2) for cane yield. Environments in blue – Genotypes in red.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

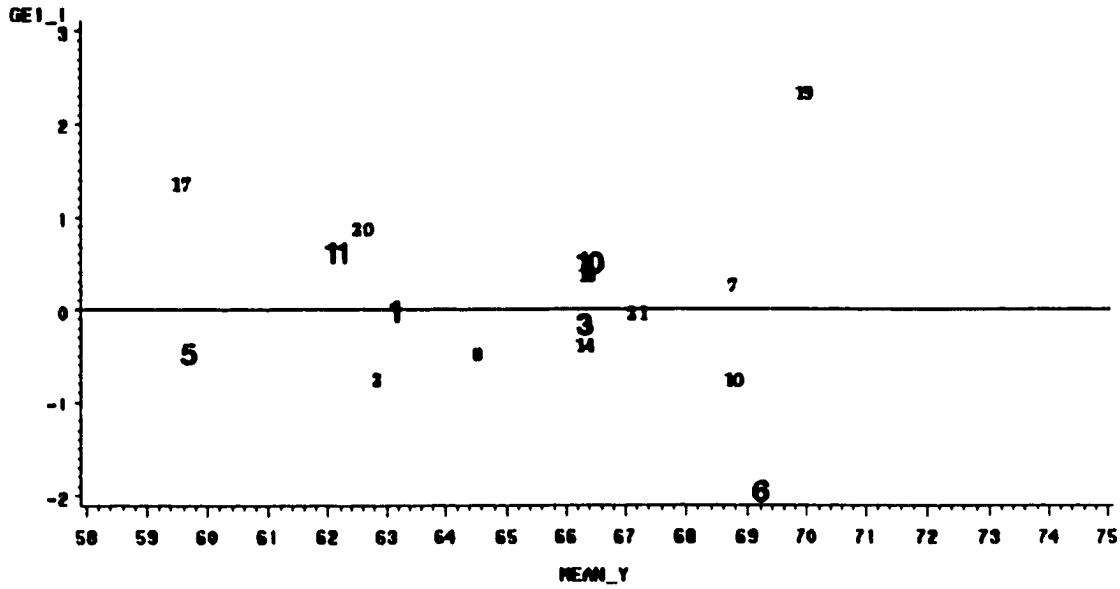

Figure 4.2. Genotype and environment scores on first multiplicative term of a Mixed AMMI vs. cane yield means [Mg ha^{-1}]. Environments in blue – Genotypes in red.

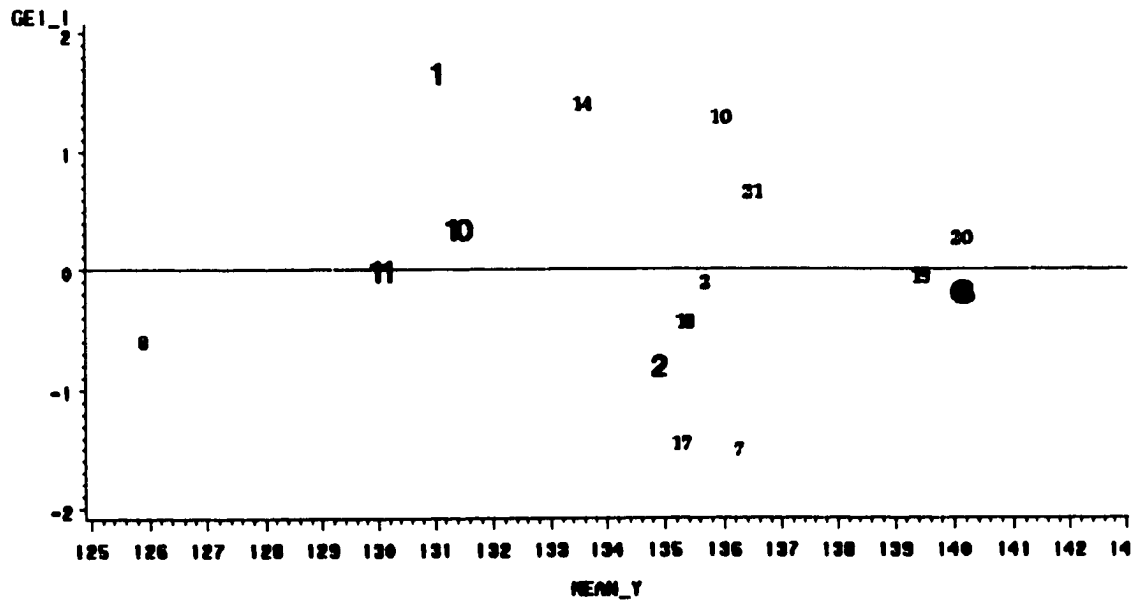

Figure 4.3. Genotype and environment scores in the multiplicative term of a Mixed AMMI (1) vs. sucrose content [g sucrose kg^{-1} cane] for year 1998.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

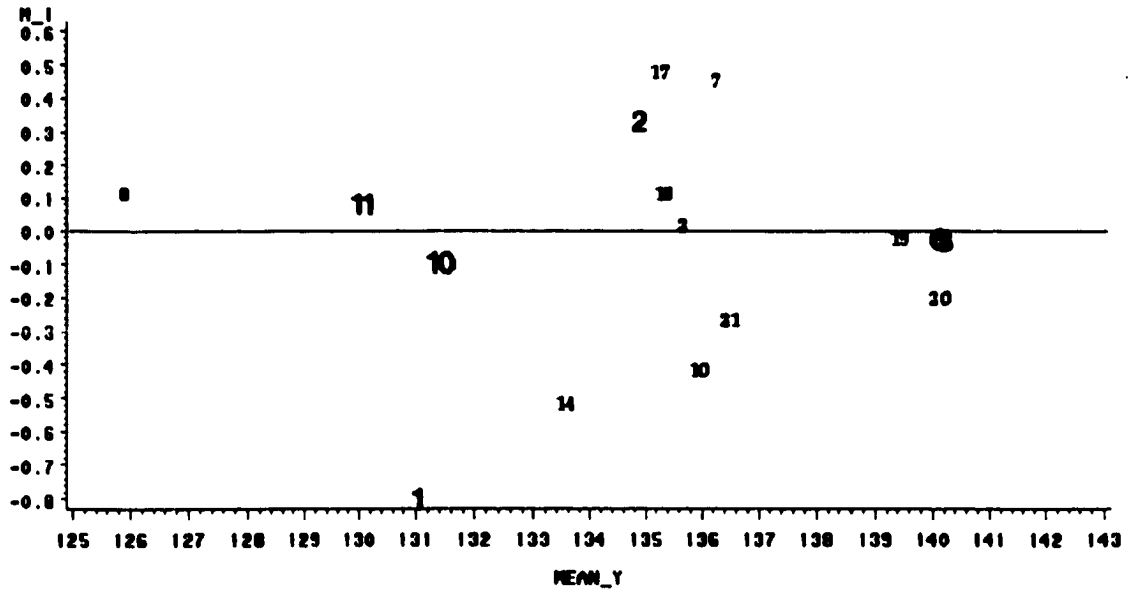

Figure 4.4. Genotype and environment scores in the multiplicative term of a FIXED AMMI (1) vs. sucrose content [g sucrose kg^{-1} cane] for year 1998

The results from the crossvalidation procedure carried out to evaluate narrow inference showed that the fixed model approach consistently produced larger prediction errors than the other models (Table 4.9). On average, the fixed model approach produced errors of 11.406 Mg ha^{-1} compared with the mixed models mean values between 9.738 and 10.399 Mg ha^{-1}. The model with the lowest root mean square error, however, varied by year and location. The mixed AMMI model [3] produced the lowest errors in nine out of 23 location-year combinations. However, the other two mixed models, the mixed ANOVA [1] and the mixed Shukla [2], each produced the lowest errors in seven out of 23 tests. The mixed E&R [4] model showed larger prediction error for this data set. The modeling of heterogeneous residual variances [5] produced larger improvement over model [1] that the models involving heterogeneous GE. This may be related with the small GE interaction observed in the LSVDP outfield trials. Such variability among the different predictive models was also observed for sucrose content

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(data not shown). Even though, the empirical nature of the prediction errors reported here, the observed trends in the simultaneous comparison of the models support one to prefer a mixed model instead of a fixed model. The researcher should use a likelihood ratio test when deciding on what model approach to use for a given analysis. To facilitate comparison of the mixed models to the fixed model, prediction errors were obtained for only test years involving a particular location. The fixed approach (genotype-environment mean) can be applied only in environments with data. However, by using a mixed model approach, predictions can also be done for genotype-environment combinations not actually evaluated.

Mixed model narrow inferences incorporate expected GE effects. P(50|50) values were used to assess the functional effect of the models in cane yield rankings by variety (Table 4.10). As with the conclusions drawn from the root mean square prediction errors, the mixed approaches were generally better than the fixed approach. Adjusting for inter-trial residual variability was important. The best model varied with year and location.

Although the mixed ANOVA [5] and [2] models were on average better than the other models, the AMMI model [3] most often gave the best predictor. The model with heterogeneous by environment residual variance showed significant improvement. One should note that these are results based on typically 10 genotypes within a location. Simultaneous comparison of these models with regard to narrow inferences in METs involving a larger genotype list will provide more insight.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.9. Prediction error† for environment-specific inferences for six models to analyze cane yield in multi-environment sugarcane yield trials for four traits.

Farm	Year Test	Fixed Model	Mixed ANOVA [1]	Mixed Shukla [2]	Mixed AMMI [3]	Mixed E&R [4]	Mixed HetR [5]
		---------- Mg ha^{-1} ----------					
1:ALL.	1998	4.079	3.746	3.918	3.660	3.981	3.830
	1996	4.707	3.920	3.730	3.901	3.912	4.065
2:B.S.	1998	13.550	12.609	13.344	12.130	12.591	12.180
	1997	7.083	5.728	5.837	6.158	6.044	5.669
	1996	11.420	9.610	8.305	9.872	10.261	10.080
3:Geo	1998	6.511	5.744	6.220	6.028	6.041	5.900
	1997	10.508	10.500	10.163	9.702	10.197	10.243
4:Gln.	1997	3.700	3.700	3.458	2.886	3.524	3.284
	1996	17.081	13.766	15.65	11.968	14.635	13.880
5:Lan.	1998	20.552	17.090	18.520	21.488	23.422	18.720
	1997	13.611	10.260	9.940	10.024	10.489	9.940
	1996	7.633	7.170	6.540	7.370	7.551	7.543
6:Mag.	1998	24.665	23.340	24.50	23.05	24.092	24.940
	1997	6.432	6.267	6.470	6.612	6.930	6.094
	1996	11.165	9.374	8.400	8.300	9.290	9.193
7:Oak.	1996	8.655	7.990	5.620	8.570	9.062	7.903
8:P.A.	1997	10.283	8.661	8.863	9.549	9.795	8.351
	1996	7.092	5.566	5.140	5.043	5.633	5.723
9:R.L.	1996	7.327	6.441	5.468	7.010	6.689	5.307
10:R.H	1998	13.176	10.947	11.482	11.460	11.420	10.234
.	1997	21.183	18.483	18.751	17.813	20.22	17.623
	1996	14.478	12.130	10.162	13.192	13.179	12.380
11:St.J.	1998	17.458	12.167	13.490	13.260	13.567	11.978
Mean		11.406	9.792	9.738	9.959	10.399	9.785

† Square root of the mean square prediction error (difference between predictor and target values for each genotype obtained by an iterative validation procedure).

† 1:All.-A.V.Allain & Sons, 2:B.S.-Bon Secour, 3:Geo.-Georgia, 4:Gln.-Glenwood, 5: Lan.-Lanaux, 6: Mag.-Magnolia, 7:Oak.-Oaklawnhy, 8: P.A.-Palo Alto, 9:R.L.-Raceland, 10: R.H.-Ronal Hebert, 11:StJ.-Levert-St. John.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4.10. Percent of top 50% varieties in a particular environment also in the top 50% of the environment specific variety performances predicted by six statistical models

Farm	Year	Fixed Model	Mixed ANOVA	Mixed Shukla	Mixed AMMI	Mixed E&R	Mixed HetR
		----------	----------	----------	% ----------	----------	----------
1:ALL.	1998	78	80	84	86	80	85
	1996	79	84	85	84	83	84
2:B.S.	1998	62	67	66	73	67	71
	1997	78	79	80	80	80	80
	1996	59	64	62	62	60	64
3:Geo	1998	53	54	53	53	52	53
	1997	77	77	77	80	77	80
4:Gln.	1997	75	75	75	81	73	78
	1996	69	73	71	74	72	71
5:Lan.	1998	67	70	67	64	65	68
	1997	72	72	72	70	72	75
	1996	66	81	75	68	75	78
6:Mag.	1998	68	73	64	74	68	70
	1997	89	98	98	97	97	98
	1996	54	67	70	64	69	69
7:Oak.	1996	74	83	84	77	83	85
8:P.A.	1997	56	68	67	72	65	70
	1996	78	78	79	79	77	77
9:R.L.	1996	79	73	79	75	75	75
10:R.H	1998	53	41	42	43	40	42
.	1997	85	78	77	70	77	80
	1996	67	64	69	65	68	64
11:St.J.	1998	54	60	55	62	60	64
Mean		69.2	72.1	71.8	71.9	71.6	73.1

† 1:All.-A.V.Allain & Sons, 2:B.S.-Bon Secour, 3:Geo.-Georgia, 4:Gln.-Glenwood, 5: Lan.-Lanaux, 6: Mag.-Magnolia, 7:Oak.-Oaklawnhy, 8: P.A.-Palo Alto, 9:R.L.-Raceland, 10: R.H.-Ronal Hebert, 11:StJ.-Levert-St. John.

4. CONCLUSIONS

The use of mixed models to analyze advanced variety trials offers the potential to improve predictive precision at virtually no additional cost. It also enables the researcher to objectively incorporate GE stability measures with mean performance. More complex mixed models that may model within environment covariance or

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

consider environmental factors linked to GE are possible (Biarnes-Dumoulin et al.,

1996; Cullis, et al., 1997; Magari and Kang, 1997; Wolfinger and Tobias, 1998).

5. REFERENCES

Becker,H.C., Leon, J. 1988. Stability analysis in plant breeding. Plant Breeding 101:1-23.

Biarnes-Dumoulin, V.,Denis J.B.,Lejeune-Henaut, Eteve G.. 1996. Interpreting yield stability in pea using genotype and environmental covariates. Crop Sci. 36:115-120.

Crossa, J.1990. Statistical analyses of multilocation trials. Advances in Agronomy 44: 55-85.

Cullis, B.R., Thompson, F.M., Fisher, J.A., Gilmour, A.R., Thompson, R. 1996. The analysis of the NSW wheat variety database. II. Variance component estimation. Theor. Appl. Genet. 92:28-39.

Denis, J.B., Gower, J.C .1996. Asymptotic confidence regions for biaditive models: interpreting genotype-environment interactions. Appl Statistics 45:479-493

Eberhart, SA Russell, W.A. 1966. Stability parameteres for comparing varieties. Crop Sci 6:36-40

Gabriel, K.R. 1971. Biplot display of multivariate matrices with application to principal components analysis. Biometrika 58: 453-467

Gauch, H.G. Jr.1988. Model selection and validation for yield trials with interaction. Biometrics 44:705-715.

Guillot, D.P., Milligan, S.B., Bischoff, K.P., Quebedeaux, K.L, Gravois, K.A., Garrison, D.D., Jackson W.R., Waguespack, H.L.1998. 1998 Outfield variety trials. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 88-101.

Gollob, H.F.1968. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrica 33:73-115.

Hill Jr., R.R., Rosenberg, J.L. 1985. Models for combining data from germplasm evaluation trials. Crop Sci 25:467-470

Jenrich, R.L., Schluchter, M.D. 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42:805-820.

Kang, M.S., Miller, J.D. 1984. Genotype x environment interactions for cane and sugar yield and their implications in sugar breeding. Crop Sci. 24:435-440.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Kang, M.S. (ed.) 1990. Genotype-by-environment interaction and plant breeding. Dep. of Agronomy, Louisiana State Univ. Agric. Center, Baton Rouge, LA.

Kang, M.S. , Gauch, H.G.,Jr. (ed.).1996. Genotype by environment interaction. CRC Press, Boca Raton, Fl.

Kang M.S. , Magari, R. 1996. New developments in selecting for phenotypic stability in crop breeding. In M.S. Kang and H.G. Gauch, Jr (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Lin, C.S.,Binns M.R., Lefkovitch L.P. 1986. Stability analysis. Where do we stand? Crop Sci. 26:894-900

Lin, C.S.,Binns M.R.1994. Concepts and methods for analyzing regional trial data for cultivar and location selection. Plant Breed. Rev. 12, 271-297

Littell, R.C., Milliken, G.A., Stroup ,W.W., Wolfinger, R.D. 1996. SAS® System for Mixed Models. Cary, N.C.:SAS Institute Inc.

Mandel , J. 1971. A new analysis of variance model for non-additive data. Technometrics 13:1-18

McLean, R.A., Sanders, W.L., Stroup, W.W. 1991. A unified approach to mixed linear models. American Statistician 45:54-64.

Oman, S.D. 1991. Multiplicative effects in mixed models analysis of variance. Biometrika 78 729:739

Piepho, H.P. 1994. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects multiplicative interaction (AMMI) analysis. Theor Appl Genet 89:647-654

Piepho, H.P.1997. Analyzing genotype-environment data by mixed models with multiplicative effects. Biometrics 53:761-766

Piepho, H.P. 1998a. Stability analysis using the SAS System. Agron. J.

Piepho, H.P. 1998b. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor. Appl. Genet. 97:195-201.

Price, W.J. , Shafii, B. 1993. The use of biplots in diagnosing interaction patterns of two-way classification data. Proceedings of the Eighteenth Annual SAS Users Group International Conference, Cary (NC): SAS Institute, Inc.

Shafii, B., Price ,W.J. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. Journal Agric., Biol. and Environ Stat. 3:335-345.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Quebedeaux, K.L., Milligan, S.B., Martin, F.A, Garrison, D.D., Jackson W.R., Waguespack, H. Jr. 1996. 1996 Outfield variety trials. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 77-98.

Quebedeaux, K.L., Milligan, S.B., Martin, F.A, Garrison, D.D., Jackson W.R., Waguespack, H. Jr. 1997. 1997 Outfield variety trials. Sugarcane Res. Annual Progress Report. Louisiana State University Agricultural Center. Louisiana Ag. Exp. Stn., 70-90.

SAS Institute. 1997. SAS/STAT software: changes and enhancements through release 6.12. SAS Inst., Cary, NC.

Searle, S.R., Casella ,G., McCulloch, C.H. 1992. Variance Components. Wiley, New York.

Shukla, G.K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237-245.

Stroup, W.W., Mulitze, D.K. 1991. Nearest neighbor adjust best linear unbiased prediction. Am. Stat. 45:194-200

Van Eeuwijk, F.A.,Denis J.B., Kang, M.S. 1996. Incorporating additional information on genotype and environments in models for two-way genotype by environment tables. In M.S. Kang and H.G. Gauch, Jr. (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Vargas M., Crossa J.,Sayre K., Reynolds M., Ramirez M.E., Talbot M. 1998. Interpreting genotype by environment interaction in wheat by partial least square regression. Crop Sci. 38:679-689.

Weber, W.E., Wricke, G., Westerman, T. Selection of genotypes and prediction of performance by analyzing genotype-by-environment interactions. In M.S. Kang and H.G. Gauch, Jr. (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL

Williams, E.J.1952. The interpretation in factorial experiments. Biometrika 39:65-81.

Wolfinger, R.D., Tobias R., 1998. Joint estimation of location, dispersion, and random effects in robust design. Technometrics, 40(1):62-71

Yates F, Cochran W.G.1938. The analysis of groups of experiments. J.Agric Sci, Cambridge, 28:556-580.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Yau,S.K. 1995. Regression and AMMI analysis of genotype by environment interactions. An empirical comparison. Agron. J. 87:121-126

Zobel W.R., Wright M.J., Gauch H.G.. 1988. Statistical analysis of a yield trial. Agron. J. 80:388-393

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 5

CONCLUSIONS AND FINAL REMARKS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Best linear unbiased prediction is an important tool for plant breeders that can be employed at several stages of the selection process. The BLUP of genetic effects may substitute cross and genotype means in progeny tests and early selection stages. In progeny tests and early selection stages, the random nature of genotypes supports the use of mixed models. A large number of genotypes facilitates estimation of genetic variance components and random effects. In later selection stages, genotypes may be assumed as fixed effects. By assuming environments and genotype-environment terms as random, variances and covariances may be modeled and hence more information can be integrated into broad and narrow genotype inferences and GE analysis.

BLUP prediction is not a new technique. What is relatively new for plant breeders, since software for handling general mixed model has become available, is the possibility of easily defining BLUPs of random effects that contemplate the model complexity and the size of databases in typical crop improvement programs. A single important "BLUP" does not exist in plant breeding. There are a large number of different combinations of fixed and random effects that can be predicted. For each value to be predicted, there are many alternative models differing with regard to the variance-covariance structure of the random effects.

Even when the main interest is in the fixed effects, parsimonious models of the covariance structure increase the prediction accuracy of performance predictors. Mixed model approaches can integrate genotype-by-environment covariances into the comparison of the genotype means. Mixed models in yield trials in several environments unifies under one general procedure the estimation of stability parameters,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the study of GE and yield mean performances. Mixed AMMI models and corresponding biplots to visualize predictable GE patterns should be obtained from specific procedures, but their interpretation is analogous to that from fixed AMMI models. Assumptions of balanced data are not made, but normality is required for maximum likelihood estimation procedures to be performed. This later point may be a limitation. There, however, exists an important amount of phenotypic information where these procedures can be applied. Other limitations to the use of mixed model analysis are related to computer time and the possible lack of convergence of likelihood-based algorithms employed to estimate variance components. Both problems may be tackled by adjusting the number of model parameters to be simultaneously estimated. Usually there exists more than one strategy to fit the same model. Working with mean values instead replications may be a functional alternative.

BLUP-based cross predictions consistently improved, with respect to MPV prediction, the accuracy of predicted performance of crosses that have never been made or tested. Different versions of BLUP could be obtained depending on the procedure selected to connect tested and untested cross effects and the model for random genetic effects. A mixed linear model adjusting progeny test data for fixed trial effects and partitioning the genotype effect into random female and male effects performed better than the model that used random cross effects. Results indicated that there was no gain by using information from "related by pedigree" crosses. This failure was attributed to the low genetic variance components and to the crossing techniques involved in sugarcane breeding.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

BLUPs to combine genotype effects from early sugarcane clonal stages demonstrated improvement compared to the percent of the check methods. BLUPs of genotype-environment combinations in yield trials also performed better than the mean to predict genotype performance in a particular environment. Progeny testing has proved to be effective and cost efficient for sugarcane breeding because it improves the efficiency of early generation selection. It can also be exploited to generate BLUPs of untested crosses and to choose parental germplasm to combine in new hybrids. The MP-BLUP obtained from the databases of regular sugarcane progeny tests would facilitate the identification of material to be crossed. Predictors based on progeny tests can be obtained after one year of testing of new parents, whereas parental selection based only on clonal *per se* information requires several years of testing and probably is more biased by the presence of non-additive genetic effects. No additional field experiments are required to calculate cross performance predictors in those programs that are already performing progeny tests. When using genotype BLUPs for prediction of future selection stages, BLUPs and SP were superior to the predictor APCH used by the LSDVP. In addition, BLUP accuracy was not dependent on check values, thus they can still be effectively used when check varieties fail.

Better performance predictions increased the probability of selecting the best genotypes at crossing, early and late selection stages of the breeding program. The improved prediction methodology may enable the breeders to increase the selection intensity at earlier stages and possibly shorten selection cycles.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX A

MODELS TO PREDICT GENOTYPE PERFORMANCE

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note: All codes assume that a SAS data set named '**FULLT**' is available with variables **Y**, **GENO**, **TRIAL** and **REP** related to the trait values y_{ijk} and codings for genotypes, trials involving at least one of the **GENO** for which a performance prediction is required, and replicates, respectively. **FULLT** contains all the data regarding trials that involve the varieties of interest, the file should contain a variable **CTXPARS** whit value equal to 1 for the varieties of interest and equal to zero otherwise. The variety **TRIAL** could be a combination of year, locations, series, etc. For description of the models used to obtain performance predictors see Table 3.1, pag. 57.

MODELS [1] TO [4] : Check-based predictors

```
/* DEFINE THE VARIETIES THAT SHOULD BE USED AS CHECKS  */
DATA CHECKID;
INPUT GENO @@;
CHECK='YES';
CARDS;
65357 70321 74383 72370
82089 83153 85384 85845

PROC SORT DATA=FULLT;BY TRIAL;
PROC MEANS NOPRINT;BY TRIAL;
VAR Y;
OUTPUT OUT=TRIALID N=NTRIAL;

DATA _NULL_;
SET TRIALID END=EOF;
CALL SYMPUT('GROUP' ||LEFT(_N_),TRIM(TRIAL));
IF EOF THEN CALL SYMPUT ('TOTAL', _N_);
RUN;

DATA _NULL_;
SET CHECKID END=EOF;
CALL SYMPUT('CHECK'||LEFT(_N_),TRIM(GENO));
IF EOF THEN CALL SYMPUT ('NCHECK', _N_);
RUN;

%MACRO PREDICT;

DATA WORKDS;
SET FULLT;
%DO C=1 %TO &NCHECK;

    DATA CHECK&C;
    SET WORKDS;
    IF TRIM(GENO)="&&CHECK&C";
    Y&C=Y;
    KEEP TRIAL REP Y&C;
    PROC SORT DATA=CHECK&C;BY TRIAL REP;
    PROC SORT DATA=WORKDS;BY TRIAL REP;
    DATA WORKDS;
    MERGE CHECK&C WORKDS;BY TRIAL REP;
    PCTCH&C=100*(Y/Y&C);

%END;

    DATA APCTCH;
    SET WORKDS;
    IF CTXPARS=1;
    APCTCH=MEAN(OF PCTCH1 PCTCH2 PCTCH3 PCTCH4
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
                PCTCH5 PCTCH6 PCTCH7 PCTCH8);
    PROC SORT DATA=CHECKID;BY GENO;
    PROC SORT DATA=FULLT;BY GENO;
    DATA ACHECK;
    MERGE FULLT CHECKID;BY GENO;
    IF CHECK='YES';

    PROC SORT DATA=ACHECK ;BY TRIAL;
    PROC MEANS DATA=ACHECK MEAN NOPRINT; BY TRIAL;
    VAR Y;
    OUTPUT OUT=ACHECKT MEAN=YCHECKT STDERR=SCHECKT;

    DATA VARLST;
    MERGE ACHECKT APCTCH;BY TRIAL;
    YPCT_EC=100*(Y/YCHECKT);
    Z=(Y-YCHECKT)/SCHECKT;
    DIFF=Y-YCHECKT;

    PROC SORT DATA=VARLST; BY GENO;
    PROC MEANS MEAN NOPRINT;BY GENO;
    VAR  Y APCTCH YPCT_EC Z DIFF;
    OUTPUT OUT=FIXED MEAN=YM1 YM2 YM3 YM4 YM5;
    PROC RANK DATA=FIXED OUT=RANKF;VAR YM1 YM2 YM3 YM4 YM5 _TYPE_;
    RANKS M1RANK M2RANK M3RANK M4RANK M5RANK MT;

    PROC MEANS DATA=ACHECK MEAN NOPRINT;
    VAR Y;
    OUTPUT OUT=MUCHECK MEAN=MUCHECK STD=SCHECK;
    DATA PREDICT;
    MERGE RANKF MUCHECK;BY _TYPE_;
    KEEP GENO YM1 YM2 YM3 YM4 YM5 P1 P2 P3 P4 P5
    M1RANK M2RANK M3RANK M4RANK M5RANK;
    P1=YM1;
    P2=(YM2*MUCHECK)/100;
    P3=(YM3*MUCHECK)/100;
    P4=(YM4*SCHECK)+MUCHECK;
    P5=YM5+MUCHECK;
    PROC PRINT;

%MEND PREDICT;
%PREDICT
```

MODEL [5]: LSMean from a Fixed Model as predictor

```
%MACRO PREDICT;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC MIXED DATA=WORKDS NOITPRINT NOCLPRINT METHOD=REML;
CLASS TRIAL GENO REP;
MODEL Y=TRIAL REP(TRIAL) GENO;
LSMEANS GENO;
MAKE 'LSMEANS' NOPRINT OUT=LSMEAN;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE LSMEAN VARLST;
IF CTXPARS=1;
P1=_LSMEAN_; DROP VARIETY;

DATA PREDICT;
SET PREDICT;
VARIETY=CLON;

PROC RANK DATA=PREDICT OUT=RANKF;VAR P1;
RANKS M1RANK;

%MEND PREDICT;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

MODEL [6]: BLUP of genotype effect -BGa-

```
%MACRO PREDICT;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC MIXED DATA=WORKDS noitprint noclprint method=REML;
CLASS TRIAL GENO REP;
MODEL Y=TRIAL REP(TRIAL)/SOLUTION;
RANDOM GENO/S;
MAKE 'SOLUTIONR' NOPRINT OUT=BLUP;
MAKE 'SOLUTIONF' NOPRINT OUT=MU;

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_;
DROP GENO;

DATA PREDICT;
MERGE PREDICT MU;BY CTXPARS;
GENO=CLON;
P1=MU+U1;

PROC PRINT;
PROC RANK DATA=PREDICT OUT=RANKF;VAR P1;
RANKS M1RANK;

%MEND PREDICT;
```

MODEL [7]: BLUP of genotype effect -BGb-

```
%MACRO PREDICT;

DATA FULLT;
SET FULLT;
IF  GENO=65357 OR GENO=70321
    OR GENO=74383 OR GENO=72370 OR GENO=82089  THEN NEW=0;
    ELSE NEW=1;
    IF (NEW) THEN GENTYPE=999999;
    ELSE GENTYPE=GENO;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC MIXED DATA=WORKDS noitprint noclprint method=REML;
CLASS TRIAL GENO REP gentype;
MODEL Y=TRIAL REP(TRIAL) gentype/SOLUTION;
RANDOM GENO*new/S;
lsmeans gentype;
MAKE 'SOLUTIONR' noprint OUT=BLUP;
MAKE 'SOLUTIONF' NOPRINT OUT=MU;

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
ID CTXPARS NEW;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_;
DROP GENO;

DATA PREDICT;
MERGE PREDICT MU;BY CTXPARS;
GENO=CLON;
P1=MU+U1;
IF NEW=1;

%MEND PREDICT;
```

MODEL [8]: BLUP of genotype effect -BGTa-

```
%MACRO PREDICT;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC MIXED DATA=WORKDS NOITPRINT NOCLPRINT METHOD=REML;
CLASS TRIAL GENO REP;
MODEL Y=/SOLUTION;
RANDOM TRIAL REP(TRIAL)  GENO/S;
MAKE 'SOLUTIONR' NOPRINT OUT=BLUP;
MAKE 'SOLUTIONF' NOPRINT OUT=MU;

DATA BLUP;
SET BLUP; IF _EFFECT_='GENO';

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_; DROP GENO;

DATA PREDICT;
MERGE PREDICT MU;BY CTXPARS;
VARIETY=CLON;P1=MU+U1;PROC PRINT;

PROC RANK DATA=PREDICT OUT=RANKF;VAR P1;
RANKS M1RANK;

%MEND PREDICT;
```

MODEL [9]: BLUP of genotype effect -BGTb-

```
%MACRO PREDICT;

DATA FULLT;
SET FULLT;
IF  GENO=65357 OR GENO=70321
    OR GENO=74383 OR GENO=72370 OR GENO=82089  THEN NEW=0;
    ELSE NEW=1;
    IF (NEW) THEN GENTYPE=999999;
    ELSE GENTYPE=GENO;

DATA WORKDS;
SET FULLT;
CLON=GENO;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
PROC MIXED DATA=WORKDS NOITPRINT NOCLPRINT METHOD=REML;
CLASS TRIAL GENO REP GENTYPE;
MODEL Y= GENTYPE/SOLUTION;
RANDOM TRIAL REP(TRIAL) GENO*NEW/S;
LSMEANS GENTYPE;
MAKE 'SOLUTIONR' NOPRINT OUT=BLUP;
MAKE 'SOLUTIONF' NOPRINT OUT=MU;

DATA BLUP;
SET BLUP;
IF _EFFECT_='NEW*GENO';

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS new;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_;
DROP GENO;

DATA PREDICT;
MERGE PREDICT MU;BY CTXPARS;
GENO=CLON;
P1=MU+U1;
IF NEW=1;

%MEND PREDICT;
```

MODEL [10]: BLUP of genotype effect -BGTIa-

```
%MACRO PREDICT;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC SORT;BY TRIAL GENO ;
PROC MEANS DATA=WORKDS MEAN NOPRINT;BY TRIAL GENO;
VAR Y;
ID CTXPARS TRIAL;
OUTPUT OUT=YMEAN MEAN=;

PROC MIXED DATA=YMEAN NOITPRINT NOCLPRINT METHOD=REML;
CLASS TRIAL GENO;
MODEL Y=TRIAL/SOLUTION;
RANDOM INT/SUBJECT=GENO S;
REPEATED GENO/SUBJECT=TRIAL  R;
MAKE 'SOLUTIONR'  OUT=BLUP;
MAKE 'SOLUTIONF'  OUT=MU;

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_;DROP GENO;

DATA PREDICT;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
MERGE PREDICT MU;BY CTXPARS;
GENO=CLON;
P1=MU+U1;

PROC PRINT;
PROC RANK DATA=PREDICT OUT=RANKF;VAR P1;
RANKS M1RANK;

%MEND PREDICT;
```

MODEL [11]: BLUP of genotype effect -BGTIb-

```
%MACRO PREDICT;
****ESTIMATE WEIGHTS ****;
PROC SORT DATA=FULLT;BY TRIAL GENO;
PROC MEANS MEAN NOPRINT;BY TRIAL GENO;
VAR Y;
ID CTXPARS;
OUTPUT OUT=YMEAN MEAN=;

PROC MIXED DATA=YMEAN NOITPRINT NOCLPRINT METHOD=REML;
CLASS TRIAL GENO;
MODEL Y=TRIAL/SOLUTION;
RANDOM INT/SUBJECT=GENO S;
REPEATED TRIAL/SUBJECT=GENO  R;
MAKE 'COVPARMS' OUT=VC;

DATA VC_R;
EST=1;
OUTPUT;
DATA VC;
SET VC VC_R;
PROC SORT DATA= FULLT;BY TRIAL;

PROC GLM  DATA= FULLT OUTSTAT=FIRSTS; BY TRIAL;
CLASS REP GENO;
MODEL Y=REP GENO;

DATA WEIGHT1;
SET FIRSTS;
IF _SOURCE_='REP';
REP=DF;
KEEP TRIAL REP;

DATA WEIGHT2;
SET FIRSTS;
IF _SOURCE_='ERROR';
SIGMA2=SS/DF;
KEEP TRIAL SIGMA2;

DATA WEIGHT;
MERGE WEIGHT1 WEIGHT2;
BY TRIAL;
IF DF=0 THEN DO;SIGMA2=10;REP=1;END;
WT=1/(SIGMA2/REP);

PROC SORT DATA=SOURCE1.FULLT;BY TRIAL;
DATA FULLT;
MERGE FULLT WEIGHT;BY TRIAL;

DATA WORKDS;
SET FULLT;
CLON=GENO;

PROC SORT;BY TRIAL GENO ;
PROC MEANS DATA=WORKDS MEAN NOPRINT;BY TRIAL GENO;
VAR Y WT;
ID CTXPARS;
OUTPUT OUT=YMEAN MEAN=Y WT;

PROC MIXED DATA=YMEAN  METHOD=REML;
CLASS TRIAL GENO;
WEIGHT WT;
MODEL Y=TRIAL/SOLUTION;
RANDOM INT/SUBJECT=GENO S;
RANDOM TRIAL/SUBJECT=GENO;
REPEATED;
PARMS/PDATA=SOURCE1.VC EQCONS=3;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
MAKE 'SOLUTIONR' NOPRINT  OUT=BLUP;
MAKE 'SOLUTIONF' NOPRINT  OUT=MU;


DATA BLUP;
SET BLUP;
IF _EFFECT_='INTERCEPT';

DATA MU;
SET MU;
IF _EFFECT_='INTERCEPT';
MU=_EST_;KEEP MU CTXPARS;
CTXPARS=1;

PROC SORT DATA=WORKDS;BY CLON;
PROC SUMMARY NWAY;
CLASS CLON;
ID CTXPARS;
OUTPUT OUT=VARLST;

DATA PREDICT;
MERGE BLUP VARLST;
IF CTXPARS=1;
U1=_EST_;
DROP GENO;

DATA PREDICT;
MERGE PREDICT MU;BY CTXPARS;
GENO=CLON;
P1=MU+U1;

PROC PRINT;
PROC RANK DATA=PREDICT OUT=RANKF;VAR P1;
RANKS M1RANK;

%MEND PREDICT;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX B

SAS CODES FOR MIXED MODELS IN MULTI-ENVIRONMENT YIELD TRIALS.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note: All codes assume that a SAS data set named '**YLD**' is available with variables **Y**, **GENO**, **ENV** and **REP** related to the trait values y_{ijk} and codings for genotypes, environments and replicates, respectively. The file *pdmixmac.sas* contains a SAS macro named PDMIXMAC612 which show mean separation results from a mixed model fitting by using letter groups. It can be obtained from the SAS web page.

MODEL [1] : Mixed ANOVA
Fixed genotypes, random environmental and replication(environment) effects. Assumptions for the GE terms: homogeneous variance and independence

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM ENV REP(ENV) GENO*ENV;
LSMEANS GENO/PDIFF;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;

%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```

MODEL [2] : Mixed Shukla
Fixed genotypes, random environmental and replication(environment) effects. Assumptions for the GE terms: Heterogeneous by genotypes variances and independence

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM INT REP/SUBJECT=ENV;
RANDOM GENO/SUBJECT=ENV TYPE=UN(1);
LSMEANS GENO/PDIFF;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;

%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```

MODEL [3a] : Mixed AMMI(1)
Fixed genotypes, random environmental and replication(environment) effects. Assumptions for the GE terms: Heterogeneous by genotypes variances and covariances between GE terms of two genotypes in the same environment.

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM INT REP/SUBJECT=ENV;
RANDOM GENO/SUBJECT=ENV TYPE=FA0(1);
LSMEANS GENO/PDIFF;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;
%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

MODEL [3b] : Mixed AMMI(2)

Fixed genotypes, random environmental and replication(environment) effects. Assumptions for the GE terms: Heterogeneous by genotypes variances and covariances between GE terms of two genotypes in the same environment.

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM INT REP/SUBJECT=ENV;
RANDOM GENO/SUBJECT=ENV TYPE=FA0(2);
LSMEANS GENO/PDIFF;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;
%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```

MODEL [4] : Mixed E&R

Fixed genotypes, random replication(environment) effects. Assumptions for the GE terms: Heterogeneous by genotypes variances and covariances between GE terms of two genotypes in the same environment.

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM REP/SUBJECT=ENV;
RANDOM GENO/SUBJECT=ENV TYPE=FA1(1);
LSMEANS GENO/PDIFF;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;
%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```

MODEL [5] : Mixed HetR

Fixed genotypes, random environmental and replication(environment) effects. Assumptions for the GE terms: homogeneous variance and independence Assumptions for the error tems: heterogeneous by environment variance.

```
PROC MIXED DATA=YLD COVTEST;
CLASS ENV GENO REP;
MODEL Y=GENO;
RANDOM ENV REP(ENV) GENO*ENV;
LSMEANS GENO/PDIFF;
REPEATED/GROUP=ENV;
MAKE 'DIFFS'   OUT=P NOPRINT;
MAKE 'LSMEANS' OUT=M NOPRINT;
%INCLUDE 'A:PDMIXMAC.SAS';
%PDMIX612(P,M,ALPHA=.05,SORT=YES);
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX C

CROSSVALIDATION IN MULTI-ENVIRONMENT TRIALS INVOLVING RANDOMIZED COMPLETE BLOCK DESIGNS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note: The macro RCBD_CV input variables are the number of runs for the validation procedure (**NCHECK**) and the data set name (**DATA**) containing the variables **Y**, **GENO**, **ENV** and **REP** related to the trait values y_{ijk} and codings for genotypes, environments and replicates, respectively.

The macro RCBD_CV calls another macro (RUNMIX) which is attached to the end of the RCBD_CV macro. The macro RUNMIX allows to run several models for multi-environment trial at each run of the cross-validation procedure. RUNMIX needs the file pdmixmac.sas. The file *pdmixmac.sas* contains a SAS macro named PDMIXMAC612 which show mean separation results from a mixed model fitting by using letter groups. It can be obtained from the SAS web page.

```
%MACRO RCBD_CV (NCHECK=100,DATA=YLD);

        PROC SORT DATA=&DATA;BY ENV;
        PROC MEANS MEAN NOPRINT;BY ENV;
        VAR Y;
        OUTPUT OUT=OUTENV MEAN=;

        PROC SORT DATA=&DATA;BY REP;
        PROC MEANS MEAN NOPRINT;BY REP;
        VAR Y;
        OUTPUT OUT=OUTREP MEAN=;

        DATA _NULL_;
        SET OUTENV END=EOF;
        CALL SYMPUT ('ENV'||LEFT(_N_),ENV);
        IF EOF THEN CALL SYMPUT('NENV',_N_);

        DATA _NULL_;
        SET OUTREP END=EOF;
        IF EOF THEN CALL SYMPUT('NREP',_N_);

        DATA CHECKDS;
           %DO K=1 %TO &NCHECK;
              %DO I=1 %TO &NENV;
                CHECK=&K;
                ENV="&&ENV&I";
                REP_OUT=CEIL(RANUNI(100)*&NREP);
                OUTPUT;
              %END;
            %END;


/* CREATING BASE FILES FOR APPENDING RESULTS*/;
        DATA A.NI;
        LENGTH STRUCTR $20;
        LENGTH GENO    $12;
        LENGTH MSGROUP $20;
        LENGTH ENV     $20;
        STRUCTR=' ';VARIETY='';ENV='';_SE_=.;
        _PRED_=.;CHECK=.; Y=.; SPE=.;MSGROUP=' ';

     %DO K=1 %TO &NCHECK;
        DATA SAMPLE;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
        SET CHECKDS;
        IF CHECK=&K;
        PROC SORT DATA=SAMPLE;BY ENV;
        PROC SORT DATA=&DATA;BY ENV REP;

        DATA WORKDS;
        MERGE &DATA SAMPLE;BY ENV;
        IF REP NE REP_OUT;
        %LET DSNAME=WORKDS;

        DATA VALIDS;
        MERGE &DATA SAMPLE;BY ENV;
        IF REP=REP_OUT;

        DATA ESTIM;
        LENGTH STRUCTR $20;
        LENGTH GENO    $12;
        LENGTH ENV     $12;
        LENGTH MSGROUP $20;
        STRUCTR=' ';GENO='';ENV=' ';
         _SE_=.;_PRED_=.; MSGROUP=' ';


        /* FITING MODELS */;

        *MIXED ANOVA;
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM ENV REP(ENV) GENO*ENV,
        COMMENT=2W);

        *MIXED SHUKLA;
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM INT REP/SUB=ENV ;
        RANDOM GENO/SUB=ENV TYPE=UN(1),
        COMMENT=SV);

        *MIXED E&R;
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM REP/SUB=ENV ;
        RANDOM GENO/SUB=ENV TYPE=FA1(1),
        COMMENT=FW, OUTLSM=ESTIM);

        *MIXED AMMI(2);
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM  INT REP/SUB=ENV ;
        RANDOM GENO/SUB=ENV TYPE=FA0(2),
        COMMENT=ER);

        *MIXED ANOVA WITH HETEROGENEOUS RESIDUAL VARIANCE;
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM ENV REP(ENV) GENO*ENV;
        REPEATED/GROUP=ENV;
        COMMENT=H2W);

        *FIXED MODEL;
        %RUNMIX(METHOD=METHOD=REML,
        Z=RANDOM REP,
        COMMENT=2WF);

        /*START CALCULATION OF PREDICTION ERRORS*/;
        PROC SORT DATA=VALIDS;BY ENV GENO;

        PROC SORT DATA=ESTIM;BY ENV VARIETY;

        DATA SPCI&K;
        MERGE VALIDS ESTIM;BY ENV VARIETY;
        CHECK=&K;
        KEEP ENV VARIETY _PRED_ Y  STRUCTR CHECK SPE MSGROUP _SE_;
        SPE= (Y-_PRED_)*(Y-_PRED_);

        PROC SORT;BY CHECK ENV STRUCTR _PRED_;
        PROC APPEND BASE=A.NI DATA=SPCI&K FORCE;

%END;
%MEND;
%RCBD_CV;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

/*MACRO RUNMIX */

```
%MACRO RUNMIX(METHOD=,Z=,COMMENT=);

    %LET _PRINT_=OFF;

    %IF (&COMMENT NE 2WF) %THEN %DO;
      PROC MIXED DATA=&DSNAME &METHOD;
      ID GENO ENV;
      CLASS GENO ENV REP;
      MODEL Y=GENO/P;
      &Z;
      LSMEAN GENO/PDIFF;
      MAKE 'PREDICTED' OUT=PRD&COMMENT NOPRINT;
      PROC SORT DATA=PRD&COMMENT;BY ENV GENO;
      PROC MEANS DATA=PRD&COMMENT NOPRINT;BY ENV GENO;
      VAR _PRED_ _SEPRED_;
      OUTPUT OUT=ESTIMS MEAN=;
      DATA ESTIMS;
      SET ESTIMS;
      KEEP STRUCTR _PRED_ MSGROUP ENV GENO _SE_;
      _SE_=_SEPRED_;
      MSGROUP=' ';
      STRUCTR="&COMMENT";
      PROC APPEND BASE=ESTIM DATA=ESTIMS FORCE;

   %END;


   %IF (&COMMENT=2WF) %THEN %DO;
      PROC SORT DATA=&DSNAME;BY ENV;
      PROC MIXED DATA=&DSNAME &METHOD;BY ENV;
      CLASS GENO  REP ;
      MODEL Y=GENO;
      &Z;
      LSMEANS GENO/PDIFF;
      MAKE 'DIFFS' OUT=P&COMMENT NOPRINT;
      MAKE 'LSMEANS' OUT=M&COMMENT NOPRINT;
       PROC PRINT DATA=M&COMMENT;
       PROC PRINT DATA=P&COMMENT;
      %INCLUDE 'A:PDMIXMAC.SAS';
      %PDMIX612(P&COMMENT,M&COMMENT,ALPHA=.05,SORT=YES);
      DATA ESTIMS;
      SET MSGRP;
      STRUCTR="&COMMENT";
      _PRED_=_LSMEAN_;
      KEEP STRUCTR _PRED_ MSGROUP ENV VARIETY _SE_;
      STRUCTR="&COMMENT";
      PROC APPEND BASE=ESTIM DATA=ESTIMS FORCE;
    %END;

    %LET _PRINT_=ON;
%MEND RUNMIX;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

APPENDIX D

BIPLOTS FOR MIXED AMMI MODELS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note: This program fit Mixed AMMI with one and two multiplicative terms and produce Biplots to visualize results.

It assumes that there exists a file data set named **YLD** containing the variables related to the trait values y_{ijk} and codings for genotypes, environments and replicates, respectively. Genotype and environment should assume numeric values.

```
GOPTIONS CBACK=WHITE;
OPTIONS NOCENTER LS=75;
LIBNAME A 'A:';
/*************READ IN DATA*******/
%LET ENV_N = 7;   /*** SET THE NUMBER OF ENVIRONMENTS ***/
%LET GEN_N= 11;   /*** SET THE NUMBER OF GENOTYPES    ***/
%LET REP_N= 3;    /*** SET THE NUMBER OF REPLICATES   ***/
%LET VAR = Y;      /*** SET THE NAME OF THE RESPONSE VARIABLE     ***/
%LET ENV = ENV;    /*** SET THE NAME OF THE ENVIRONMENT VARIABLE  ***/
%LET GEN = GEN;    /*** SET THE NAME OF THE GENOTYPE VARIABLE     ***/
%LET REP = REP;    /*** SET THE NAME OF THE REPLICATION VARIABLE  ***/

/********FITTING MIXED AMMI*********************/
%MACRO MIXED;
PROC MIXED DATA=YLD;
CLASS  &ENV &GEN &REP;
MODEL &VAR = &GEN/P PM;
RANDOM INT &REP/SUBJECT=ENV;
RANDOM &GEN/SUBJECT=ENV TYPE=FA0(1) S;
MAKE 'PREDICTED' OUT=A.PRED1FR NOPRINT;
MAKE 'PREDMEANS' OUT=A.PRED1F  NOPRINT;
MAKE 'SOLUTIONR' OUT=A.GBYE1;
MAKE 'COVPARMS'  OUT=A.COV1;
MAKE 'FITTING'   OUT=A.FA1;
ID &GEN &ENV &REP;


DATA GBYE;
SET A.GBYE1;
IF _EFFECT_='GEN';
GE1=_EST_;
KEEP  GE1;
PROC SORT DATA=YLD;
BY &ENV &GEN;
PROC MEANS DATA=YLD NOPRINT;
BY &ENV &GEN;
VAR &VAR;
OUTPUT OUT=MEANS MEAN=YIJ;
DATA VEC;
MERGE MEANS GBYE;
KEEP &ENV &GEN GE1 YIJ;


PROC MIXED DATA=YLD;
CLASS &ENV &GEN &REP;
MODEL &VAR = &GEN/P PM;
RANDOM INT &REP/SUBJECT=ENV;
RANDOM &GEN/SUBJECT=ENV TYPE=FA0(2) S;
MAKE 'PREDICTED' OUT=A.PRED2FR NOPRINT;
MAKE 'PREDMEANS' OUT=A.PRED2F  NOPRINT;
MAKE 'SOLUTIONR' OUT=A.GBYE2;
MAKE 'COVPARMS'  OUT=A.COV2;
MAKE 'FITTING'   OUT=A.FA2;
ID &GEN &ENV &REP;

DATA GBYE;
SET A.GBYE2;
IF _EFFECT_='GEN';
GE2=_EST_;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
KEEP GE2;
DATA A.VEC;
MERGE VEC GBYE;
KEEP &ENV &GEN GE1 GE2 YIJ;

%MEND MIXED;
%MIXED;

DATA VECG1;
SET A.COV1;KEEP GEN_1;
IF SUBSTR(COVPARM,6,1)='1' OR SUBSTR(COVPARM,7,1)='1';
GEN_1=EST;
KEEP GEN_1;

DATA VECG21;
SET A.COV2;
IF SUBSTR(COVPARM,6,1)='1' OR SUBSTR(COVPARM,7,1)='1';
GEN_21=EST;
KEEP GEN_21;

DATA VECG22;
SET A.COV2;
IF SUBSTR(COVPARM,4,1)='2' AND SUBSTR(COVPARM,6,1)='1' THEN FLAG=1;
IF SUBSTR(COVPARM,6,1)='2' OR  SUBSTR(COVPARM,7,1)='2' THEN FLAG=1 ;
IF FLAG=1;
GEN_22=EST;
KEEP GEN_22;

DATA VECG;
MERGE VECG1 VECG21 VECG22;
GEN_1=-1*GEN_1;
GEN_21=-1*GEN_21;
GEN_22=-1*GEN_22;
GEN+1;
KEEP GEN_1 GEN_21 GEN_22 GEN;

PROC SORT DATA=A.VEC;
BY ENV GEN;


PROC IML;
USE VECG;
READ ALL INTO GLOAD;
K2=GLOAD[,2:3];
COEFF=I(&ENV_N)@K2;
USE A.VEC;
READ ALL INTO GBYE;
GBYE2=GBYE[,5];
ELOAD2=GINV(COEFF)*GBYE2;
PRINT ELOAD2;
K1=GLOAD[,1];
COEFF=I(&ENV_N)@K1;
GBYE1=GBYE[,4];
ELOAD1=GINV(COEFF)*GBYE1;
PRINT ELOAD1;
GSCORE=K1||K2;
ELOAD2M=SHAPE(ELOAD2,7);

VECE=ELOAD1||ELOAD2M;
CREATE GSCORE FROM GSCORE [COLNAME={GEN_1 GEN_21 GEN_22}];
APPEND FROM GSCORE;
CREATE VECE FROM VECE [COLNAME={WI1 WI21 WI22}];
APPEND FROM VECE;

/*standardization of the genotype scores*/
PROC MEANS DATA=GSCORE MEAN STD NOPRINT;
VAR GEN_1 GEN_21 GEN_22 ;
OUTPUT OUT=SALG ;
PROC PRINT DATA=SALG;

DATA NEWVEC;
SET GSCORE;
GEN1_Z=(GEN_1+0.3619)/1.1371;   /*USE MEAN AND STANDARD DEVIATION OF GEN_1*/
GEN21_Z=(GEN_21+0.2775)/1.1346; /*USE MEAN AND STANDARD DEVIATION OF GEN_1*/
GEN22_Z=(GEN_22-0.4428)/1.1275; /*USE MEAN AND STANDARD DEVIATION OF GEN_1*/

PROC SORT DATA=YLD;
    BY &ENV;

PROC MEANS DATA=YLD NOPRINT;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
    VAR &VAR;
    BY &ENV;
    OUTPUT OUT=ENVA MEAN=YDOT;

PROC SORT DATA=YLD;
    BY &GEN;

PROC MEANS DATA=YLD NOPRINT;
    VAR &VAR;
    BY &GEN;
    OUTPUT OUT=GENA MEAN=YDOT;

DATA ENVA;
    MERGE ENVA VECE;

DATA GENA;
    MERGE GENA NEWVEC; KEEP &GEN YDOT GEN21_Z GEN22_Z GEN1_Z;
IF YDOT=. THEN DELETE;
PROC PRINT DATA=GENA;

/****************BIPLOT 1: Two fisrt multiplicative terms ******************/

DATA ENVANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET ENVA;
    TEXT=&ENV;
    STYLE = 'SWISSB';
    XSYS='2'; YSYS='2'; COLOR='BLUE'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1.5;
    X=WI21;
    Y=WI22;

DATA GENANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET GENA;
    TEXT=&GEN;
    STYLE = 'ZAPFB';
    XSYS='2'; YSYS='2'; COLOR='RED'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1;
    X=GEN21_Z;
    Y=GEN22_Z;

DATA VECANN1;
    SET ENVANNO GENANNO;

DATA VECTORS;
    SET ENVA GENA;

PROC GPLOT DATA=VECTORS;
    SYMBOL1 V=NONE I=NONE COLOR=WHITE;
    PLOT GEN22_Z*GEN21_Z=1 WI22*WI21=1/ANNO=VECANN1 OVERLAY VREF=0 HREF=0;
    TITLE1 'MIXED AMMI(2). FIRST AND SECOND MULTIPLICATIVE TERM;
    TITLE2 'GENOTYPE NUMBER IN RED - ENVIRONMENT NUMBER IN BLUE;
RUN;

/*****************Biplot 2. First multiplicative term vs yield mean **************/

DATA ENVANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET ENVA;
    TEXT=&ENV;
    STYLE = 'SWISSB';
    XSYS='2'; YSYS='2'; COLOR='BLUE'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1.5;
    X=YDOT;
    Y=WI21;

DATA GENANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET GENA;
    TEXT=&GEN;
    STYLE = 'ZAPFB';
    XSYS='2'; YSYS='2'; COLOR='RED'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1;
    X=YDOT;
    Y=GEN21_Z;


DATA VECANN2;
    SET ENVANNO GENANNO;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
DATA VECTORS;
    SET ENVA GENA;

PROC GPLOT DATA=VECTORS;
    SYMBOL1 V=NONE I=NONE COLOR=WHITE;
    PLOT GEN21_Z*YDOT=1 WI21*YDOT=1/ANNO=VECANN2 OVERLAY VREF=0;
    TITLE1 'MIXED AMMI(2) FIRST MULTIPLICATIVE TERM VS. YIELD MEAN';


/*****************SECOND MULTIPLICATIVE TERM VS YIELD MEAN ******************/
DATA ENVANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET ENVA;
    TEXT=&ENV;
    STYLE = 'SWISSB';
    XSYS='2'; YSYS='2'; COLOR='BLUE'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1.5;
    X=YDOT;
    Y=WI22;


DATA GENANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET GENA;
    TEXT=&GEN;
    STYLE = 'ZAPFB';
    XSYS='2'; YSYS='2'; COLOR='RED'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1;
    X=YDOT;
    Y=GEN22_Z;

DATA VECANN3;
    SET ENVANNO GENANNO;

DATA VECTORS;
    SET ENVA GENA;

PROC GPLOT DATA=VECTORS;
    SYMBOL1 V=NONE I=NONE COLOR=WHITE;
    PLOT GEN22_Z*YDOT=1 WI22*YDOT=1/ANNO=VECANN3 OVERLAY VREF=0;
    TITLE1 'MIXED AMMI(2) SECOND MULTIPLICATIVE TERM VS. YIELD MEAN';

RUN;

/*****FIRST MULTIPLICATIVE TERM VS YIELD MEAN FOR AN AMMI(1) ******************/

DATA ENVANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET ENVA;
    TEXT=&ENV;
    STYLE = 'SWISSB';
    XSYS='2'; YSYS='2'; COLOR='BLUE'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1.5;
    X=YDOT;
    Y=WI1;


DATA GENANNO(KEEP=XSYS YSYS X Y COLOR FUNCTION POSITION SIZE TEXT STYLE);
    LENGTH TEXT $ 8;
    SET GENA;
    TEXT=&GEN;
    STYLE = 'ZAPFB';
    XSYS='2'; YSYS='2'; COLOR='RED'; POSITION='5'; FUNCTION='LABEL';
    SIZE=1;
    X=YDOT;
    Y=GEN1_Z;

DATA VECANN2;
    SET ENVANNO GENANNO;

DATA VECTORS;
    SET ENVA GENA;

PROC GPLOT DATA=VECTORS;
    SYMBOL1 V=NONE I=NONE COLOR=WHITE;
    PLOT GEN1_Z*YDOT=1 WI1*YDOT=1/ANNO=VECANN2 OVERLAY VREF=0;
    TITLE1 'MIXED AMMI(1) FIRST MULTIPLICATIVE TERM VS. YIELD MEAN';

RUN;
```


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

VITA

Mónica Balzarini was born on February 9, 1962 in Córdoba, Argentina. She graduated with a bachelor of science degree in Agronomy from the Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias, in 1984. In 1985, she was appointed for working in the Biometry Unit of that institution. In the same year, she was awarded with a research grant from the Research Council of Córdoba. She taught courses in Agricultural Experimentation and Statistics, and offered statistical consulting for agricultural problems. She was awarded a fellowship for Master Science level training from the University of Córdoba. In 1995, she received a M.S. degree in Biometry from Universidad de Buenos Aires and Instituto Nacional de Tecnología Agropecuaria, Argentina. After that, she participated in several national and international research projects and supervised agricultural scientist research work. In 1995, she received a University award because of her teaching and research activities in applied statistics from the University of Córdoba.

In 1996, she obtained a grant from the Argentinean Government (FOMEC) to visit the Department of Experimental Statistics, Louisiana State University, United States, and in 1997 she was awarded a graduate assistantship from the Department of Agronomy, Louisiana State University to support her doctoral studies. She obtained her doctor of philosophy degree in 2000. She concentrated her studies in statistics and quantitative genetics. She was nominated to Sigma Xi, The Scientific Research Society, Louisiana Chapter. She is now supervising a group of scientists working on statistics applied to the agricultural sciences in the Universidad Nacional de Córdoba, Argentina.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Monica Graciela Balzarini

Major Field: Agronomy

Title of Dissertation: Biometrical Models for Predicting Future Performance in Plant Breeding

Approved:

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

Date of Examination:

March 16, 2000

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.