APPLICATIONS OF BEST LINEAR PREDICTION TO THE ANALYSIS OF FIVE FULL-SIB LOBLOLLY PINE PROGENY TESTS

by

T. La Farge

Abstract.--Two applications of Best Linear Prediction (BLP) were demonstrated and compared. One method obtains variance and covariance components (homogeneous second moments) from a combined ANOVA based on theory and expected values to predict breeding values for all parents in all environments in a breeding zone. The other method combines variances of family means from separate ANOVAS from each test with family mean correlations between tests to obtain heterogeneous second moments. When there are large environmental differences and genotype x environment interactions such that there are distinct breeding zones, the second method may be used to predict the breeding values of those parents best suited for specific target environments.

Keywords: Pinus taeda L., Best Linear Prediction, full-sib progeny test, type B family mean correlation, breeding value.

INTRODUCTION

Anyone who has worked in a large progeny testing program recognizes the need for appropriate, powerful and robust systems for analyzing progeny test data and estimating breeding values for roguing seed orchards and making selections for second- and future-generation seed orchards. Although well established statistical methods exist for analyzing such data for a wide range of mating and experimental designs, these traditional methods assume balanced data sets at all levels of replication. When these equalities are seriously violated, traditional methods do not work well without costly and time-consuming adjustments, such as missing plot estimates.

Animal breeders have long been acutely aware of this problem, since the data sets they work with rarely have equality of replication. Balanced data sets are almost unachievable, since different herds are not equal in size, and it is not practical to equalize them. Therefore, animal breeders have been compelled to seek methods in which selection could be accomplished in very unbalanced data (Henderson 1984, White and Hodge 1989). The two most

1/ Eastern Zone Geneticist, Region 8, U.S.D.A. Forest Service, Atlanta, Georgia.

recently developed methods are Best Linear Prediction (BLP) and Best Linear Unbiased Prediction (BLUP). The utility of both methods is currently growing due to the increasing availability and power of computers, especially of personal computers. This paper will be concerned only with BLP and its analysis on a Compaq 386 personal computer, since BLUP requires a more powerful computer system for any large data set.

METHODS AND MATERIALS

The Essential Components of the BLP Equation

The methods used in these analyses are described by White et al.(1986) for half-sib tests and by White and Hodge (1989) for many other applications, including full-sib progeny tests. The latter source should be consulted for any questions concerning the theory of BLP. A principal utility of BLP is that it can be used to evaluate the performance of parental genotypes in a full-sib progeny test.

A principal assumption of BLP is that the first and second moments are known (Henderson 1984, White et al. 1986). Second moments are specified in two matrices, C and V. The C matrix is a nonsymmetric matrix which defines the genetic relationships between the observed full-sib family means at each site and the true but unknown breeding values, g. Each column of C represents a parental breeding value to be predicted. The elements comprising C are calculated from genetic theory (White et al. 1986).

The V matrix is a symmetric matrix which represents the variances and covariances between the observed phenotypic values. The main diagonal comprises variances of family means for each planting location. In full-sib progeny tests the covariances in the off-diagonals that are not zero are covariances between family means which refer to either: (1) different tests with two common parents; (2) different tests with one common parent; or (3) the same test with one common parent (White and Hodge 1989). Once these matrices have been specified, breeding values can be predicted by means of the following formula:

$$\hat{g} = C'V^{-1}y \qquad (1)$$

where C and V are defined above, y = a data vector of observed deviations of the family means at each location from the location mean, and a = the breeding values to be predicted.

The Data Analyzed

The trait considered in the following analyses was total height at age 5, and observations were full-sib family means at each site. All data were collected from four loblolly pine (Pinus taeda L.) progeny tests in southern and one in northern Mississippi. Test locations were as follows:

Test Number	Ranger District	National Forest	Number of Full-sib Families	Number of Check Lots
070001	Black Creek	Desoto	20	4
070002	Strong River	Bienville	26	4
070004	Homochitto	Homochitto	28	4
070005	Strong River	Bienville	28	4
070006	Holly Springs	Holly Springs	29	5

Tests 070001 and 070002 were planted in 1978, and tests 070004, 070005 and 070006 were planted in 1979.

The numbers of families plus check lots that any two tests had in common ranged from a low of 14 to as many as 32. All families comprised seven unrelated 6 x 6 element diallel crossing groups, none of which were complete. The number of crosses per crossing group ranged from two to 13 of the possible 15. Also, tests 070001 and 070002 had four replications, whereas tests 070004, 070005 and 070006 each had three replicates. Hence, considerable imbalance existed in the overall data set.

Two Approaches to BLP

Basically, there are two ways of approaching BLP. One method obtains variance and covariance components (homogeneous second moments) from a combined analysis of variance based on theory and expected values. The other method combines variances of family means from separate ANOVAS for each test with family mean correlations between tests (Type B correlations, Burdon 1977) to obtain heterogeneous second moments. The first method assumes equal variances at all sites and equal levels of genotype x environment interactions between all pairs of sites; the second method is useful if the different sites have unequal variances and assumes different levels of interactions between different pairs of sites. We may obtain breeding values for a target environment which differs in some respect, say elevation, from the other environments being sampled. Both methods are evaluated and compared in the following discussion.

Homogeneous Second Moments

To implement the first method, the five tests were analyzed as a combined ANOVA for total height by means of the VARCOMP Procedure of the Statistical Analysis System (SAS 1987) for Personal Computers. This analysis obtained the variance components needed to derive appropriate variances and covariances to be entered into the C and V matrices included in the BLP equation.

A full-sib family mean of the cross of parents j and k at a single site i is designated Y_{ijk}. The variance of family means from the combined ANOVA of the test was $Var\ (y_{ijk}) = 1.0088$ with $b = 3.3333$ replicates per test and $n = 8.9738$ trees per plot, harmonic mean basis. The variance of family means comprises all elements on the main diagonal of the V matrix. The covariance of family means with two common parents in separate tests was $Cov\ (y_{ijkl}, y_{i'jkl}) = 0.6243$. The covariance among families with one

common parent in separate tests was $Cov(y_{ijkl}, y_{i',jkl}) = 0.2930$. The covariance among families in the same test-with one common parent was $Cov(y_{ijkl}, y_{ijkl'}) 0.3337$. These covariances comprise the off diagonal element of the V matrix. The covariance between the observed family mean and the true breeding value was $Cov(y_{ijkl}, g) = 0.5860$. This last covariance is the single constant element comprising the C matrix.

The next step was to use PROC MEANS to obtain means for each family at each location. These data were modified by obtaining the differences between the family means and each location mean. These marginal means form the y vector shown in equation 1.

All of the above matrices and vectors may be manually loaded into the format appropriate for running SAS Interactive Matrix Language (IML), but for large data sets it was essential to devise a runstream which loaded these matrices automatically. The language needed to write such programs is defined in the SAS IML Guide for Personal Computers (1985). Such programs were used to load all matrices used in these analyses. Portions of the V and C matrices used in these analyses are presented in Tables 1 and 2.

Table 1. A portion of the V matrix derived from the combined ANOVA of the loblolly pine full-sib progeny test.

V=11.0088	.6243	.6243	.3337	.3337	.3337	.3337	.3337	.3337	.3337
.6243	1.0088	.6243	.3337	.3337	.3337	.3337	.3337	.3337	.3337
.6243	.6243	1.0088	.3337	.3337	.3337	.3337	.3337	.3337	.3337
.3337	.6243	.6243	1.0088	.2930	.2930	.3337	.3337	.3337	.3337
.3337	.3337	.3337	.3337	.3337	.6243	.6243	.6243	.6243	1.00881;

Since the V matrix shown in Table 1 represents a single diallel crossing group, it does not indicate the full size and complexity of the full data set. The largest matrix was a 50 x 50 element matrix, the smallest a 10 x 10 element matrix. A full matrix comprising all family x location means would have required a 152 x 152 element matrix. The largest that our Compaq 386 PC could invert in the IML procedure was a 65 x 65 element matrix. The subdivision into seven unrelated crossing groups made it possible to construct matrix subsets with no loss of information.

Heterogeneous Second Moments

The procedure utilizing heterogeneous second moments requires an additional step. In this example separate sets of breeding values were predicted, each set targeted for the environment represented by one of the test locations sampled. First, ANOVAs were performed to obtain variances of family means for each location. Second, two kinds of correlations were obtained between all locations: (1) correlations based on two common parents (full-sib family means); and (2) correlations based on one common parent (half-sib family means). The variances of family means and correlations were then combined to obtain family mean covariances (Type B family mean covariances, Burdon 1977; White and Hodge 1989) among all test locations according to the following general formula:

$$Cov(y_{ijk}, y_{i'jk}) = r_{Bf} \left[Var(y_{ijk}) \, Var(y_{i'jk}) \right]^{1/2} \quad (2)$$

where $Cov(y_{ijk}, y_{i'jk})$ is the Type B covariance of family means, r_{Bf} is the correlation between family means in each test, and $Var(y_{ijk})$ and $Var(y_{i'jk})$ are the variances of family means of each of the two tests respectively.

Table 2. A portion of the C matrix derived from the combined ANOVA of the loblolly pine full-sib progeny test.

C={.5860	0	0	0	.5860	0,
.5860	0	0	0	.5860	0,
.5860	0	0	0	.5860	0,
.5860	0	0	.5860	0	0,
0	0	0	.5860	0	0,
0	0	0	0	0	.5860};

A subset of the V matrix comprising some of the heterogeneous variances and covariances produced by this approach is presented in Table 3, and a subset of the C matrix comprising the appropriate covariances is presented in Table 4.

Table 3. A portion of the V matrix derived from separate ANOVAs of loblolly pine full-sib progeny tests and Type B correlations and covariances.

V={1.0485	.7600	.5282	.4171	.5110	1.0019	.5430	.4155	1.0019	.5430
.7600	1.3308	.9553	.4830	.2216	.5430	.8021	.7709	.5430	.8021
.5282	.9553	1.5112	.6765	.0442	.4155	.7709	1.4217	.4155	.7709
.4177	.4830	.6765	1.1012	.5419	.7748	.9626	.9796	.4171	.4830

Table 4. A portion of the C matrix derived from separate ANOVAS and Type B correlations and covariances.

C={.8342	1.0220	2.0038	1.0282	.8310	0	0	0	0
.9660	.4432	1.0282	1.6042	1.5418	0	0	0	0
1.0492	.5366	.8342	.9660	1.3529	0	0	0	0
.5366	.8402	1.0220	.4432	.0883	0	0	0	0
.8342	1.0220	2.0038	1.0282	.8310	0	0	0	0

The following equation was used to obtain the elements of the C matrix from family mean correlations and variances of family means:

$$Cov(y_{ijk}, g) = 2 \, r_{Bf}(h_i, h_T) \left[Var(y_{ijk}) \, Var(y_{Tjk}) \right]^{1/2} \quad (3)$$

where $Cov(y_{ijk}, g)$ is the covariance between the observed family mean and the true but unknown breeding value of one parent (say k), $r_{Bf}(h_i, h_T$ test

environment and those of the target environment, and Var(y
Var(y_{Tjk}) are the variances of half-sib family means of the reference test environment and the target environment respectively.

RESULTS AND DISCUSSION

A sample of the breeding values obtained by the method using homogeneous second moments is shown in Table 5' and a sample of the breeding values for each of the five target environments is given in Table 6. Space does not permit a full listing of all breeding values.

The sample of 11 of the 38 breeding values generated by the method of homogeneous second moments shown in Table 5 has a good range and shows the general forest area (GFA) check lots to be below average. One measure of the precision of BLP is the estimated correlation between the true and predicted genetic values (CRGG):

$$CRGG = Corr(\hat{g},g) = \{Var(\hat{g})/Var(g)\}^{1/2} \quad (4)$$

(White and Hodge 1989), where $Var(\hat{g})$ is the variance of the predicted breeding values and Var(g) is the variance of the true but unknown breeding values. The CRGGs for all parents in this test equaled or exceeded 0.59, and most exceeded 0.7, which suggests very good precision for this test (homogeneous second moments only).

In southern Mississippi there is no valid reason to obtain separate sets of breeding values for each target environment. We have been unable to identify any environmental variables that provide reliable criteria for labelling such environments. For example' southern Mississippi does not have significant elevational zones. However, test 070006 is on the Holly Springs District in northern Mississippi, which is a different breeding zone. Hence, the rankings of some of the breeding values sampled in Table 6 change considerably when predicted for the target environment associated with test location 070006 on the Holly Springs District. This enables us to utilize those interactions because we can identify suitable criteria which characterize a population of such sites. If we wish to breed a set of families from southern Mississippi suitable for planting in northern Mississippi, we could set aside an orchard block consisting of those families targeted for that zone. This method of BLP based on ANOVAs in separate tests and Type B correlations and covariances between tests offers a useful and powerful tool for analyzing unbalanced data from full-sib progeny tests when target environments can be identified.

CONCLUSIONS

The precision and significance of these tests are not measurable' since we do not have the luxury of F tests or t tests. However' there are valid reasons to have some confidence in the relative precision and reliability of predicted breeding values for both methods. For example' employing the method utilizing homogeneous second moments gave high estimated correlations between the true and predicted breeding values (CRGG), on the order of 0.6 or

greater, which suggests that precise estimates of the second moments were obtained. White and Hodge (1989) note that it is probably necessary to analyze tests having at least 30 unrelated parents to obtain precise estimates of the second moments. The present set of tests would seem to satisfy this condition.

Table 5. Eleven breeding values for height at age 5 in loblolly pine for tests 070001' 070002' 070004, 070005 and 070006 obtained from best linear prediction based on homogeneous second moments.

Crossing group	Parent	Rank	Breeding value, height	Breeding value + mean height
			feet	feet
2	213	1	1.23	16.62
7	238	2	1.16	16.55
1	206	3	1.04	16.43
41	903105 1/	4	1.03	16.42
7	245	5	1.02	16.41
.	.	.	.	.
50	903413 2/	33	-0.93	14.46
8	247	34	-0.97	14.41
8	227	35	-1.23	14.16
8	243	36	-1.23	14.16
3	222	37	-1.25	14.14
3	209	38	-2 24	13.15

1/ An open-pollinated seed orchard clone. 2/ A general forest area check lot.

However' the poor sampling of the crosses resulting from the incompleteness of some of the crossing groups probably causes the predictions of those particular breeding values to be very unreliable. For example' in Table 5 parents 227 and 243 in crossing group 8 have identical breeding values because they are crossed only with each other. Decisions on the fate of parents in such poorly sampled crossing groups may be postponed until data from other tests provide more complete information.

LITERATURE CITED

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

Henderson'C.R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph' Ontario' Canada. 462 p.

SAS Institute. 1985. SAS/IML guide for personal computers' version 6 edition. SAS Institute' Inc.' Cary' North Carolina. 244 p.

SAS Institute. 1987. SAS/STAT guide for personal computers' version 6 edition. SAS Institute' Inc.' Cary' North Carolina. 1028 p.

White'T.L' G.R.Hodge and M.A.Delorenzo. 1986. Best linear prediction of breeding values in forest tree improvement. P. 99-122 in Statistical Considerations in Genetic Testing of Forest Trees' Proc. 1986 Workshop Southern Regional Information Exchange Group 40. Southern Coop. Series Bull. 324. Univ. of Florida, Gainesville' FL.

White'T.L. and G.R.Hodge. 1989 (In Press). Predicting breeding values with applications in forest tree improvement. Kluwer Academic Publishers' Dordrecht' The Netherlands.

Table 6. Six breeding values for height of loblolly pine at age 5 for each of five target environments based on data measured in tests 070001' 070002, 070004' 070005 and 070006 obtained from best linear predictions utilizing heterogeneous second moments' including rankings of each breeding value in each test.

Crossing group	Parent	Test	Rank	Breeding value' height	Breeding value + mean height
				feet	feet
2	213	070001	4	1.53	16.92
7	238	070001	1	3.16	18.55
41	903105	070001	13	0.84	16.23
5o	903413	070001	32	-2.05	13.34
8	247	070001	24	-0.76	14j 63
3	209	070001	36	-2.68	12.71
2	213	070002	6	2.06	17.45
7	238	070002	1	5.92	21.31
41	903105	070002	7	1.84	17.23
5o	903413	070002	33	-3.45	11.94
8	247	070002	32	-1.85	13.54
3	209	070002	15	0.16	15.55
2	213	070004	8	2.56	17.95
7	238	070004	1	5.21	20.60
41	903105	070004	9	1.68	17.08
5o	903413	070004	35	-4.66	10.73
8	247	070004	34	-2.14	13.25
3	209	070004	25	-1.20	14.19
2	213	070005	8	1.29	16.68
7	238	070005	2	2.41	17.80
41	903105	070005	3	2.14	17.53
5o	903413	070005	32	-1.58	13.81
8	247	070005	25	-0.57	14.82
3	209	070005	34	-2.35	13.04
2	213	070006	16	0.61	16.00
7	238	070006	31	-1.05	15.87
41	903105	070006	12	1.38	16.77
50	903413	070006	25	-0.21	15.18
8	247	070006	17	0.58	15.97
3	209	070006	37	-4.89	10.50