EXPLORATORY DATA ANALYSIS FOR A MULTIPLE PLANTATION FOREST GENETICS STUDY THROUGH A SIMPLE MEAN POLISH PROCEDURE

F. H. Kung 1/

Abstract.--Many forest genetics studies were established over a broad area with many distinct plantations. Results from such a design can be easily explored by a procedure similar to "median polish." When the mean is used instead of the median, the procedure can yield a rich collection of information. Site effects, genetic effects and genetic by site interactions are arranged neatly in a two-way table. This provides the researcher some insight to conduct a further confirmatory data analysis and is specially useful for genotype stability studies. An outline of the mean polish procedure and illustrative examples based on a white ash multiplantation progeny test are given in this paper.

The objective of exploratory data analysis is "to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it" (Tukey 1977). In this age of computers, it is all too easy and tempting to just put in the raw data from the field, run a program, and then obtain results from the general linear model. However, for a tree improvement worker, it is as important to discover as to disprove. We should look at data to see what it seems to say. The appearance of data should force a message to us and require us to look beneath it for new insights.

Most forest genetics studies are replicated in several different areas so that genetic, site, as well as genetics by site interaction variation can be tested. But if we just accept or reject a certain hypothesis with a certain probability, we are assuring ourselves of only the principle of decision-making. In practice, we need to know which seed sources should be selected and where to plant them. In this paper a simple "mean polish" procedure will display the genetic, plantation, and interaction effects neatly in a two-way table. Our mind can clearly receive the picture of various effects existing in the test. Selection for seed sources and for planting sites then can be made with ease. Height growth from a three-year-old white ash provenance test in four plantations is used in this paper as an illustration.

MODEL

From a statistician's point of view, provenance tests, progeny tests, and clonal tests can follow the same design and analysis. The difference in interpretation of variance components is due to the difference in the genetic linkage among experimental units. A multiplantation provenance test is used for illustration.

The model is $X_{ij} = M + S_i + P_j + SP_{ij}$

1/ Associate Professor, Department of Forestry, Southern Illinois University, Carbondale, Illinois 62901.

where Xij is the plantation mean of a provenance

M is the population mean

Si is the seed source effect for the ith provenance

Pj is the ith plantation effect

SPij is the interaction between seed source and plantation.

Although observations on individual trees or on plot means may be recorded in the field and may be useful for some other data analysis, the mean polish procedure presented here does not utilize information below the cell mean (i.e. plantation mean of a provenance).

For a balanced complete experimental design the seed source effect is equal to the seed source mean minus the population mean. Similarly, the plantation effect is equal to the plantation mean minus the population mean. The interaction then can be calculated by subtracting population mean, the seed source effect, and the plantation effect from the corresponding cell mean.

PROCEDURE

The procedure for mean polish can be described as follows:

1. Insert the plantation mean of a provenance in a two-way table. Use the rows for provenance and the columns for plantation.

2. Sum across each row, divide by the number of plantations, and insert the computed row mean from each cell in the original two-way table.

3. Subtract the corresponding row mean from each cell in the original two-way table.

4. Sum along each column, divide by the number of rows, and insert the computed column means in a new row.

5. Subtract the corresponding column mean from each cell in the new table (including the added column in step 2).

The r rows by c columns table in step one has been transformed into a new (r + 1) by (c + 1) table in step 5 as shown in Exhibit 1. The population mean, M, is at the lower right corner of the table, the seed source effects, are listed in the (c + 1) column, and plantation effects, Si, in the (r + 1) row. Interactions, SPij, are presented in the r by c matrix.

The procedures are similar to a two-way median fit (Gerig et al 1978) except that the means are used in place of medians. The advantage of using the mean is to obtain the least square estimate of the effects.

PRACTICAL EXAMPLE

A white ash provenance test is used here for an example. Seed was collected throughout the range in 1975 and seedlings were grown in Union State Nursery, Illinois for one year. The three-year height growth pattern in four

Exhibit 1.-- Mean polish procedure

Step	Result			
		Plt. A	Plt. B	
1. Insert cell means	Prov. P	1	5	
	Prov. Q	2	6	
	Prov. R	3	13	
2. Compute row means		1	5	3
		2	6	4
		3	13	8
3. Subtract row means from cells		-2	2	3
		-2	2	4
		-5	5	8
4. Compute column means		-2	2	3
		-2	2	4
		-5	5	8
		-3	3	5
			SP_{ij}	S_i
5. Subtract column means from each row		1	-1	-2
		1	-1	-1
		-2	2	3
		-3	3	5
			P_j	M

plantations located in Louisiana, Illinois, Ohio and Wisconsin are presented in table 1. Each plantation contains nineteen seed sources in five tree plots and five randomized complete blocks.

After the mean polish procedure was carried out, separation of effects according to the model can be seen in table 2. By examining these effects, immediately we will get the impression that plantation effect has the largest range. Hence, it is the most important factor in planting white ash. The best sites are in Ohio and Illinois. By looking at the sign of the plantation effects and with some idea of the latitude of the plantations, one can understand that a second degree regression curve would be a better choice than a simple regression line in describing the relationship between plantation height and plantation latitude.

The range of seed source effect is the smallest among the three effects. Thus, the genetic contribution would be smaller than either the plantation effect or the interaction effect. This observation can be confirmed by

variance component analysis. The plantation and the interaction contribute 72% and 13% of total variance, respectively, while the seed source contributes only 2% (Kung and Clausen 1980).

Table 1.-- Height growth of white ash at age 3

Seed Source		Outplanting in				
State	Stand	Louisiana	Illinois	Ohio	Wisconsin	Average
	No.			cm		
Me.	6785	43	70	125	37	69
Mich.	6736	43	70	141	37	73
Mich.	6779	40	70	147	37	73
W. Va.	6778	42	67	157	34	75
Conn.	6794	46	88	155	36	81
Wisc.	6723	51	75	164	45	84
Vt.	6782	50	73	187	38	87
Ala.	6733	51	135	134	33	88
La.	6738	76	153	117	13	90
Tenn.	6728	48	153	157	30	97
Tenn.	6871	50	168	143	29	97
Miss.	6737	57	173	130	30	98
Ky.	6792	34	145	183	32	98
Ky.	6734	38	167	162	29	99
Ind.	6795	35	144	180	37	99
Ill.	6721	47	155	182	34	105
Miss.	6740	63	196	131	37	107
Tx.	6768	82	184	151	24	110
Ill.	6771	52	126	214	49	111
Average		50	127	156	34	92

The range of interaction in Illinois is almost twice as large as the range of interaction in Louisiana. The favorable planting sites (Illinois and Ohio) seem to have a greater range of interaction than the unfavorable sites (Louisiana and Wisconsin). Based on this exploratory data analysis, a more rigorous F-test was conducted. It was found that the Illinois plantation had a greater interaction than the Ohio plantation, which in turn had a greater interaction than the Louisiana and the Wisconsin plantation. The difference in the size of interaction between the Louisiana and the Wisconsin plantations is not significant.

We can also look across each row in table 2 to compare the size of interaction within each seed source. Stand No. 6740 from Mississippi had the largest range of 94, while Stand No. 6721 from Illinois had the smallest range of 30. However, according to the F-test, the contrast was not significant at the 5% level due to small sample size.

The sum of square of the interaction within a genotype has been defined as "ecovalence" for that genotype (Wricke and Weber 1981). It is a measure of

Table 2.--The effect of seed source, plantation and seed source x plantation interaction on height growth of 3-year-old white ash seedlings, obtained through the mean polish procedure

Seed Source		Outplanting in				Seed Source
State	Stand	Louisiana	Illinois	Ohio	Wisconsin	Effect
	No.	Interaction Effect, cm.				
Me.	6785	16	-34	-8	27	-23
Mich.	6736	12	-28	4	22	-19
Mich.	6779	8	-38	9	21	-18
W. Va.	6778	9	-43	18	17	-17
Conn.	6794	7	-29	9	13	-10
Wisc.	6723	9	-44	16	19	-8
Vt.	6782	5	-49	36	9	-5
Ala.	6733	4	12	-18	3	-3
La.	6738	28	28	-37	-19	-2
Tenn.	6728	-8	22	-5	-9	6
Tenn.	6871	-6	35	-19	-10	6
Miss.	6737	1	40	-32	-10	6
Ky.	6792	-23	11	20	-9	7
Ky.	6734	-19	32	-1	-12	7
Ind.	6745	-22	9	17	-4	8
Ill.	6721	-15	15	13	-13	13
Miss.	6740	-1	54	-40	-12	15
Tx.	6768	13	38	-23	-28	19
Ill.	6771	-16	-20	40	-4	19
Plantation effect		-42	35	64	-58	
Population mean effect						92

genotype stability. A zero ecovalence indicates the non-existence of genotype x environment interaction. Thus, the performance of a genotype in various locations can be predicted from the mean of that genotype and the plantation effect. In contrast, a large ecovalence indicates unstable performance in various plantations. Because there were no significant differences between the largest and the smallest ecovalences, we may conclude that genotype stability was comparable among various seed sources in this study.

As forest geneticists are becoming more and more interested in the area of phenotypic stability (Morgenstern and Teich 1969), genotypic stability and adaptability (Owino and Zobel 1977), results from the simple mean polish procedure can be extremely informative. Just by comparing the direction of signs across each row with the direction of signs for the plantation effect, we can pick out those seed sources which are responsive to positive changes in environment. For example, seed sources Ky 6792, Ind 6745, and I11 6721 had the same series of signs (-, +, +, -) as the plantation effects. These seed sources can take advantage of the improved environments. They should perform better than average in better environments but less than average in poor environments. Such responsiveness is very desirable for tree improvement. As you

improve the site quality, these genotypes will give you additional height growth compared to the average tree in the average plantation. This means a higher return for your investment. To go a step further, if you want to find out which one of the three seed sources is the most sensitive, you can further run a correlation between the interaction term and the seed source effect. It turns out that seed source Ill 6721 has the highest coefficient of correlation (r = .95). Hence, that seed source is the most responsive among all provenances.

DISCUSSION

Although a multiplantation provenance test was used here as an example, the mean polish procedure can be used in other forest genetic studies which utilize a two-way cross classification design. For example, substituting provenances with clones, and plantations with fertilizer levels, we can see the genetic effect, the fertilizer effect and the interaction between them. In a controlled pollination experiment, the mean polish procedure can indicate the general combining and specific combining ability of both male and female parents.

In order to make the numbers easier to communicate, ordering the rows or columns of the table by the marginal measure of size is recommended (Ehrenberg 1981). In the given example in this paper, seed sources are sorted according to their means, and plantations are arranged according to the order of their latitudes. Such an arrangement enables us to pick up the superior seed sources and compare their performances across plantations at various latitudes. The range of seed source effects can be quickly detected. Furthermore, association between the interaction and seed source effect, as well as latitudinal trend of the interaction, can be visualized. For other experiments, the reader may find some other ordering of ecological factors, such as day length, minimum temperature or rain fall, more appropriate than latitude. In the exploratory data analysis, it is always beneficial to the researcher to try out many alternative routes to discover additional useful information.

LITERATURE CITED

Ehrenberg, A. S. C. 1981. The problem of numeracy. The American Statistician 35(2) : 67-71..

Gerig, T. M., H. T. Schreuder, D. M. Crutchfield and C. G. Wells. 1978. The two-way median fit: A sensitive statistical procedure to detect response *of* trees to fertilization. For Sci. 24(3): 358-362.

Kung, F. H. and K. E. Clausen. 1980. Height growth of white ash in four plantations. Ag Review 80, School of Agriculture, Southern Illinois University, Carbondale, Illinois, pp. 56-60.

Morgenstern, E. K. and A. H. Teich. 1969. Phenotypic stability and height growth *of* Jack pine provenances. Can. J. Genet. Cytol. 11: 110-117.

Owino, F. and B. Zobel. 1977. Genotype x environment interaction and genotypic stability in Loblolly pine. Silvae Genetica 26(l): 18-21.

Tukey, J. W. 1977. Exploratory data analysis. 506 p. Addison-Wesley Pub. Co.

Wricke, G. and W. E. Weber. 1981. Analysis of interactions in series of trials. Biometrics 37(1): 194.