GENERALISATION OF MULTI-TRAIT SELECTION INDICES USING INFORMATION FROM SEVERAL SITES

R.D. BURDON

FOREST RESEARCH INSTITUTE, NEW ZEALAND FOREST SERVICE, ROTORUA

(Received for publication 20 September 1978)

ABSTRACT

A multi-trait selection index is formulated for selecting parents using half-sib progeny test information, to illustrate the approach of treating the expressions of any one trait at different sites as being effectively several distinct traits. This involves analysing data from one site at a time and linking results from different sites in a separate but minor operation. The proposed approach is highly robust with respect to the statistical properties of data and offers great flexibility in assessment procedures and the assignment of economic weights.

INTRODUCTION

Burdon (1977) has described analysis of genotype-environment interaction, for single traits, based on regarding the expressions of any trait in different environments as effectively distinct traits. Height growth in Environment *x*, for instance, would be regarded as a different trait from height growth in Environment *y*. In this way some troublesome data characteristics can be circumvented. This general approach lends itself readily to the construction of selection indices. As an illustration this paper covers the formulation of a multi-trait selection index based on half-sib progeny test information from several sites. Applications to actual data are reported separately (Carson *et al.*, 1978; Shelbourne and Low, 1979).

THE PROGENY TEST SITUATION

Consider a half-sib progeny test uncomplicated by maternal effects, with *m* families, on each of *q* sites, with one fully randomised plot of *n* trees in each of *k* block replicates per site.

Consider analyses of variance and covariance for one site at a time. For one trait at any site assume the following analysis of variance:

Source	Degrees of freedom	Expected mean square
Families (F)	$m-1$	$\sigma^2_w + n\,\sigma^2_p + nk\,\sigma^2_f$
Replicates (R)	$k-1$	$\sigma^2_w + n\,\sigma^2_p + nm\,\sigma^2_r$
F X R (Syn. plot environment effects [within replicates])	$(m-1)(k-1)$	$\sigma^2_w + n\,\sigma^2_p$
Trees within plots	$km(n-1)$	σ^2_w

where σ^2_w, σ^2_p, σ^2_r and σ^2_f are the trees-within-plots, plot environment, replicates and families variances respectively, their estimation from the mean squares being self-evident.

The variance among family means ($\sigma^2_{\bar{f}}$) is of the form:

$$\sigma^2_{\bar{f}} = \sigma^2_f + \sigma^2_p/k + \sigma^2_w/nk \quad \ldots\ldots\ldots\ldots \quad (1)$$

Covariances between traits at a site are estimated similarly from mean cross-products in corresponding analyses of covariance.

Covariances, both for families and family means, between traits at different sites are estimated simply by the mean cross-products of family means between the traits (cf. Burdon, 1977), irrespective of whether or not the 'traits' are different in the customary sense.

THE INDEX

Adapting from Wilcox *et al.* (1975) (cf. Wilcox and Smith, 1973) the multi-trait index for selecting parents from family data is of the form:

$$I_h = \sum_{st} \hat{b}_{st}\, X^{st}_h \quad \ldots\ldots\ldots\ldots \quad (2)$$

where I_h is the index value of the h^{th} family

$\hat{b}_{st}$ is the least-squares estimate of the coefficient (weighting factor) for the t^{th} trait at the s^{th} site

X^{st}_h is the mean phenotypic value for the t^{th} trait at the s^{th} site for the h^{th} family.

The b's are estimated from:

$$\mathbf{P}\mathbf{b} = \mathbf{A}\mathbf{a}$$

$$\text{so} \quad \mathbf{b} = \mathbf{P}^{-1}\mathbf{A}\mathbf{a} \quad \ldots\ldots\ldots\ldots \quad (3)$$

where $\mathbf{P}$ is the phenotypic variance-covariance matrix for family means

$\mathbf{A}$ is the genetic variance-covariance matrix, i.e. a matrix of family variances and covariances

$\mathbf{a}$ is a column vector of economic weights assigned to a unit improvement in each trait

$\mathbf{b}$ is a column vector of weights given to the family mean for each trait.

Hence Equation 2 can be rewritten as:

$$I = \mathbf{X}'\mathbf{b} \quad \ldots\ldots\ldots\ldots \quad (4)$$

where I is the complete index

and $\mathbf{X}'$ is the row vector of family means for the various traits.

Take the case of three traits (*1, 2* and *3*) assessed at each of three sites, *x, y,* and z. The index will contain nine (i.e. 3 X 3) terms.

The elements of $\mathbf{b}$ may be listed as:

$$b_{x1},\ b_{x2},\ b_{x3},\ b_{y1},\ b_{y2},\ b_{y3},\ b_{z1},\ b_{z2},\ b_{z3}$$

where b_{x1} is the weighting factor given to family means for trait *1* at site *x* etc. The elements of $\mathbf{a}$ are strictly analogous.

The form of $\mathbf{P}$ for the given order of $\mathbf{a}$ and $\mathbf{b}$ elements, is shown in Table 1,

where $\sigma^2_{\bar{f}_{x1}}$ is the variance among family means for trait *1* at site *x*,

$\mathrm{cov}_{\bar{f}_{x1y1}}$ is the corresponding covariance of family means between trait *1* at site *x* and trait *1* at site *y*, etc. and

where $\mathbf{A}_{xy} \neq \mathbf{A}_{yx}$, etc.

The between-site submatrices are designated $\mathbf{A}_{xy}$, $\mathbf{A}_{xz}$, etc. because, as indicated earlier, in these cases the covariances of family means have the same expected values as the family covariances ($\mathrm{cov}_{\bar{f}_{x1y1}} = \mathrm{cov}_{f_{x1y1}}$, etc.).

The $\mathbf{A}$ matrix is the same form as the $\mathbf{P}$ matrix except that all the elements are family variances and covariances ($\sigma^2_{f_{x1}}$ instead of $\sigma^2_{\bar{f}_{x1}}$, etc.; $\mathrm{cov}_{f_{x1x2}}$ instead of $\mathrm{cov}_{\bar{f}_{x1x2}}$, etc.). Hence Equation 3 can be rewritten in submatrix form:

$$\mathbf{b} = \left[\begin{array}{c|c|c} \mathbf{P}_x & \mathbf{A}_{xy} & \mathbf{A}_{xz} \\ \hline \mathbf{A}_{yx} & \mathbf{P}_y & \mathbf{A}_{yz} \\ \hline \mathbf{A}_{zx} & \mathbf{A}_{zy} & \mathbf{P}_z \end{array}\right]^{-1} \left[\begin{array}{c|c|c} \mathbf{A}_x & \mathbf{A}_{xy} & \mathbf{A}_{xz} \\ \hline \mathbf{A}_{yx} & \mathbf{A}_y & \mathbf{A}_{yz} \\ \hline \mathbf{A}_{zx} & \mathbf{A}_{zy} & \mathbf{A}_z \end{array}\right] .\mathbf{a} \quad \ldots\ldots\ldots \quad (6)$$

MISSING DATA

Missing family/site subclasses cause problems, whatever form of selection index is used. In this case estimated values for the missing subclass means are needed, primarily to obtain meaningful index values for the families involved.

Work on missing data corrections (*see* Hoyle, 1971; Jarrett, 1978) has largely

TABLE 1 - Formulation of P

$$
P = \left[\begin{array}{ccc|ccc|ccc}
\sigma^2\,\bar{f}_{x1} & \mathrm{cov}\,\bar{f}_{x1x2} & \mathrm{cov}\,\bar{f}_{x1x3} & \mathrm{cov}\,\bar{f}_{x1y1} & \mathrm{cov}\,\bar{f}_{x1y2} & \mathrm{cov}\,\bar{f}_{x1y3} & \mathrm{cov}\,\bar{f}_{x1z1} & \mathrm{cov}\,\bar{f}_{x1z2} & \mathrm{cov}\,\bar{f}_{x1z3} \\
\mathrm{cov}\,\bar{f}_{x1x2} & \sigma^2\,\bar{f}_{x2} & \mathrm{cov}\,\bar{f}_{x2x3} & \mathrm{cov}\,\bar{f}_{x2y1} & \mathrm{cov}\,\bar{f}_{x2y2} & \mathrm{cov}\,\bar{f}_{x2y3} & \mathrm{cov}\,\bar{f}_{x2z1} & \mathrm{cov}\,\bar{f}_{x2z2} & \mathrm{cov}\,\bar{f}_{x2z3} \\
\mathrm{cov}\,\bar{f}_{x1x3} & \mathrm{cov}\,\bar{f}_{x2x3} & \sigma^2\,\bar{f}_{x3} & \mathrm{cov}\,\bar{f}_{x3y1} & \mathrm{cov}\,\bar{f}_{x3y2} & \mathrm{cov}\,\bar{f}_{x3y3} & \mathrm{cov}\,\bar{f}_{x3z1} & \mathrm{cov}\,\bar{f}_{x3z2} & \mathrm{cov}\,\bar{f}_{x3z3} \\
\hline
\mathrm{cov}\,\bar{f}_{x1y1} & \mathrm{cov}\,\bar{f}_{x2y1} & \mathrm{cov}\,\bar{f}_{x3y1} & \sigma^2\,\bar{f}_{y1} & \mathrm{cov}\,\bar{f}_{y1y2} & \mathrm{cov}\,\bar{f}_{y1y3} & \mathrm{cov}\,\bar{f}_{y1z1} & \mathrm{cov}\,\bar{f}_{y1z2} & \mathrm{cov}\,\bar{f}_{y1z3} \\
\mathrm{cov}\,\bar{f}_{x1y2} & \mathrm{cov}\,\bar{f}_{x2y2} & \mathrm{cov}\,\bar{f}_{x3y2} & \mathrm{cov}\,\bar{f}_{y1y2} & \sigma^2\,\bar{f}_{y2} & \mathrm{cov}\,\bar{f}_{y2y3} & \mathrm{cov}\,\bar{f}_{y2z1} & \mathrm{cov}\,\bar{f}_{y2z2} & \mathrm{cov}\,\bar{f}_{y2z3} \\
\mathrm{cov}\,\bar{f}_{x1y3} & \mathrm{cov}\,\bar{f}_{x2y3} & \mathrm{cov}\,\bar{f}_{x3y3} & \mathrm{cov}\,\bar{f}_{y1y3} & \mathrm{cov}\,\bar{f}_{y2y3} & \sigma^2\,\bar{f}_{y3} & \mathrm{cov}\,\bar{f}_{y3z1} & \mathrm{cov}\,\bar{f}_{y3z2} & \mathrm{cov}\,\bar{f}_{y3z3} \\
\hline
\mathrm{cov}\,\bar{f}_{x1z1} & \mathrm{cov}\,\bar{f}_{x2z1} & \mathrm{cov}\,\bar{f}_{x3z1} & \mathrm{cov}\,\bar{f}_{y1z1} & \mathrm{cov}\,\bar{f}_{y2z1} & \mathrm{cov}\,\bar{f}_{y3z1} & \sigma^2\,\bar{f}_{z1} & \mathrm{cov}\,\bar{f}_{z1z2} & \mathrm{cov}\,\bar{f}_{z1z3} \\
\mathrm{cov}\,\bar{f}_{x1z2} & \mathrm{cov}\,\bar{f}_{x2z2} & \mathrm{cov}\,\bar{f}_{x3z2} & \mathrm{cov}\,\bar{f}_{y1z2} & \mathrm{cov}\,\bar{f}_{y2z2} & \mathrm{cov}\,\bar{f}_{y3z2} & \mathrm{cov}\,\bar{f}_{z1z2} & \sigma^2\,\bar{f}_{z2} & \mathrm{cov}\,\bar{f}_{z2z3} \\
\mathrm{cov}\,\bar{f}_{x1z3} & \mathrm{cov}\,\bar{f}_{x2z3} & \mathrm{cov}\,\bar{f}_{x3z3} & \mathrm{cov}\,\bar{f}_{y1z3} & \mathrm{cov}\,\bar{f}_{y2z3} & \mathrm{cov}\,\bar{f}_{y3z3} & \mathrm{cov}\,\bar{f}_{z1z3} & \mathrm{cov}\,\bar{f}_{z2z3} & \sigma^2\,\bar{f}_{z3}
\end{array}\right]
= \left[\begin{array}{c|c|c}
P_x & A_{xy} & A_{xz} \\
\hline
A_{yx} & P_y & A_{yz} \\
\hline
A_{zx} & A_{zy} & P_z
\end{array}\right] \quad . \; . \; . \; (5)
$$

concentrated on insertion of values that allow analysis of variance to proceed satisfactorily rather than giving values that can be used in their own right as in selection indices. Corrections that would just give unbiased families X sites interaction mean squares are clearly inappropriate here, and they can be badly biased by between-site heterogeneity of variances among family means. A somewhat better correction would be one giving an unbiased families mean square in analysis of data from all sites, but it would be prudent to standardise (albeit approximately) the values for family means within sites.

The approach of Smith and Pfaffenberger (1970) seems appropriate for estimating values of elements in the P matrix and missing phenotypic values (of families, here), although the computations are elaborate. Estimation of values for elements in the matrix is more problematic, but the following is suggested:

$$g''_{ij} = p''_{ij} - (p_{ij} - g_{ij}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

where g''_{ij} is the required estimate for the element in the i^{th} row in the j^{th} column of the A matrix

p''_{ij} is the estimate, already obtained (*op. cit.*), of the corresponding element in the P matrix

and g_{ij} and p_{ij} are the estimates of the corresponding elements in the respective matrices that have been obtained from the incomplete data.

This estimation procedure would be superfluous for between-site covariance elements, in which $p_{ij} = g_{ij}$.

For a simple case with an isolated missing value, the following *ad hoc* solution is suggested. Consider the h^{th} family missing at site *y*. For any trait of interest at site *y* an estimated value (X''_{yh}) is needed. If the family means at site *y* are correlated better with those at site *x* than those at site *z* then the estimate will be given by

$$X''_{yh} = \bar{X}_y + b_{y.x}\,(X_{xh} - \bar{X}_x)(1 + \frac{1}{m}) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

where $\bar{X}_x$ and $\bar{X}_y$ are the observed overall averages at sites *x* and *y* respectively

$b_{y.x}$ = regression of observed family means at site *y* on means at site *x*

X_{xh} = observed mean of h^{th} family at site *x*

and $1 + \frac{1}{m}$ represents an adjustment for the fact that the estimate of $\bar{X}_y$ is computed with X_{yh} missing.

This solution gives a conservative estimate of $X_{yh} - \bar{X}_y$, which is probably desirable in the circumstances.

DISCUSSION

A more conventional way of constructing a selection index from progeny test results would be to estimate the variances and covariances from analyses that incorporate sites as a main effect. By comparison, the approach put forward has the potential disadvantage of involving a very cumbersome index because it can create a multiplicity of 'traits'*. On the credit side, there are considerable advantages (cf. Burdon, 1977):

(1) The proposed selection index does not depend at all on homogeneity, among sites, of any variance component for any trait, nor does it rely on corresponding homogeneity of between-trait covariance structures.

(2) Differences between sites even in type of experimental layout would present no major complications, while a trait would not have to be assessed on the same basis at all sites.

(3) Imbalance in the experimental classification, provided it does not involve missing family/site cells, could be handled more conveniently, because the imbalance would be isolated within the smaller classifications that exist within sites.

(4) Notably, one does not have to assess for all traits at all sites, indeed, there is no necessity for any trait to be assessed at more than one site. One might, for example, only be able to use one site for providing information on disease resistance. As another extreme case, one might just assess for growth rate at one site and just for tree form at another.

(5) Great flexibility exists in inputting economic weights. To select individuals specifically for a certain site economic weights (a's) appropriate for that site can be given to the expressions of the respective traits at that site. Zero economic weights would be given to the corresponding traits at other sites. Notwithstanding, the expressions of traits at other sites could assume worthwhile index weightings, particularly if true genotype-site interactions are minor and genotypic values are more precisely expressed at any other site. If, on the other hand, one wants to select genotypes for the range of sites one could give economic weights so as to allow both for the economic importance of each trait within a particular site and the relative importance of that site within the forest estate. Hence the final economic weight (a'_{st}) of the t^{th} trait at the s^{th} site could be of the form:

$$a'_{st} = a_{st} c_s$$

* Implausible b values can arise with essentially redundant terms in the computed index, without actually causing erroneous selections.

where a_{st} is the economic weight for the t^{th} trait within the s^{th} type

and c_s is the relative planting area represented by the s^{th} site type.

By comparing expected genetic gains from different procedures one can help decide on whether or to what degree genotypes should be selected for specific site types.

The approach can clearly be extended to using information from more than one class of relative. Consider, for example, selection of individual trees within a wind-pollinated progeny test, where there will be two classes of relatives. The full index value for an individual will contain terms that can be grouped into three classes:

(i) involving the phenotypic value of the individual, for each trait measured;

(ii) involving the individual's family mean, for each trait measured at the site where the individual was planted;

(iii) involving the individual's family mean for each trait at each other site where the trait is measured.

Standardisation of data from each site could help in ranking individual offspring from different sites. Again, there is no absolute necessity for each trait to be assessed at each site in order to get an index value, although a failure to do so would prevent any ready comparison among individuals growing at different sites.

The approach proposed here has some important advantages in flexibility, which give it greater generality in its application, although it is unlikely that any approach can be fully general in the sense of coping well with all types of data. In any case it must be remembered that least-squares index solutions can be very sensitive to errors of estimating genetic parameters and to uncertainties concerning economic weights (Namkoong, 1969; Arbez *et al.*, 1974). There can therefore be no substitute for empirical checking of genetic gain expectations in individual traits when alternative economic weights are inputted.

ACKNOWLEDGEMENTS

Thanks are extended to M.D. Wilcox, Ph. Baradat and C.J.A. Shelbourne for comments on a draft, and to I.A. Andrew for discussions on missing data corrections.

REFERENCES

ARBEZ, M., BARADAT, P., MAUGE, J.P., MILLIER, C. and BADIA, J. 1974: Some problems related to the use of selection indices in forest tree breeding. *Proc. Joint IUFRO Meeting S.02.04.01-3, Stockholm,* pp.97-116.

BURDON, R.D. 1977: Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. *Silvae Genetica 26:* 168-75.

CARSON, M.J., SHELBOURNE, C.J.A. and LOW, C.B. 1978: First assessment of uninodal progeny trials - aged 5 years from planting. *N.Z. For. Serv., For. Res. Inst., Genet. & Tree Impr. Int. Rep. 156* (unpubl.).

HOYLE, M.H. 1971: Spoilt data - an introduction and bibliography. *J. Roy. Statist. Soc. A 134:* 429-39.

JARRETT, R.G. 1978: The analysis of designed experiments with missing observations. *Appl. Statist. 27:* 38-46.

NAMKOONG, G. 1969: Problems of multiple-trait breeding. *Paper 7/4, 2nd World Consultation on Forest Tree Breeding, Washington D.C. 7-16 Aug. 1969.*

SHELBOURNE, C.J.A. and LOW, C.B. 1979: Reselection amongst 300 wind-pollinated progenies of *P. radiata* at five sites using a 17 trait combined index. *N.Z. For. Serv., For. Res. Inst., Genet. & Tree Impr. Int. Rep. 171* (unpubl.).

SMITH, W.B. and PFAFFENBERGER, R.C. 1970: Selection index estimation from partial multivariate normal data. *Biometrics 26:* 625-39.

WILCOX, M.D., FIRTH, A., LOW, C. and McCONCHIE, D. 1975: First assessment of the *Pinus radiata* open-pollinated progeny tests of the "268" series parents. Stage 2: Re-ranking of best 120 families to include density. Stage 3: Among and within family selection. *N.Z. For. Serv., For. Res. Inst., Genet. & Tree Impr. Int. Rep. 78* (unpubl.).

WILCOX, M.D. and SMITH, H.D. 1973: Selection indices for wood quality in loblolly pine. *Proc. 12th South For. Tree Impr. Conf. (Louisiana):* 322-42.