Theor Appl Genet (1990) 79:261–267

THEORETICAL AND APPLIED GENETICS

Evaluation of selection index: application to the choice of an indirect multitrait selection index for soybean breeding

A. Bouchez* and B. Goffinet

INRA, Station de Biometrie et Intelligence Artificielle, Chemin de Borde-Rouge, BP 27, F-31326 Castanet-Tolosan, France

Received July 3, 1989; Accepted November 2, 1989
Communicated by D. Van Vleck

Summary. Selection indices can be used to predict one trait from information available on several traits in order to improve the prediction accuracy. Plant or animal breeders are interested in selecting only the best individuals, and need to compare the efficiency of different trait combinations in order to choose the index ensuring the best prediction quality for individual values. As the usual tools for index evaluation do not remain unbiased in all cases, we propose a robust way of evaluation by means of an estimator of the mean-square error of prediction (EMSEP). This estimator remains valid even when parameters are not known, as usually assumed, but are estimated. EMSEP is applied to the choice of an indirect multitrait selection index at the F_5 generation of a classical breeding scheme for soybeans. Best predictions for precocity are obtained by means of indices using only part of the available information.

Key words: Selection index – Error of prediction – BLUP – Soybean

Introduction

Plant or animal breeders are interested in obtaining the most accurate possible predictions of genotypic values, in order to select the best individuals. Frequently, various observations contain information concerning the individual to be evaluated. We will consider the problem of predicting a single trait. The use of selection index allows use of more information than only the phenotypic observation of the trait to be predicted. Observations on genetically correlated traits, observations on genetically correlated individuals or individuals in the same environment can also be used. This potential information is very often unbalanced, due to either experimental disequilibrium or selection between two sets of genetically correlated individuals.

Some statistical methods allow construction of indices using all available information, regardless of the experimental or genetic conditions. They were first developed for animal breeding, and have been generalized as classical selection tools. Mixed models should fit nearly all variables (Henderson 1973) and are commonly used in animal breeding (Quaas and Pollack 1980). Henderson (1963, 1986) developed a linear index, the "best linear unbiased predictor" (BLUP), which should give the most accurate prediction in any case. Gianola (1986) has shown that under multivariate normality, BLUP, allowing for heterogeneous variances, maximizes expected genetic progress. When selection has occurred inside the observation set, Henderson (1975), Thompson (1979), and Goffinet (1983, 1987) have shown that variance components estimated by maximizing the likelihood (ML) are the only ones that lead to good prediction.

Sometimes a lot of information is available and, especially in plant breeding, one faces the question of the choice of the traits, among all the potential information, that should be used. This question is set in terms of prediction quality and expected genetic gain, but one can never ignore the cost of additional observations. To predict one single trait, plant and animal breeders look for the best overall selection strategy: direct or indirect selection, single or multitrait selection.

First, we will see how classical tools for evaluating a selection strategy fail to be reliable most of the time. The aim of this paper is to propose a robust and original method for comparing the quality of different selection

* Present address: G. S. du Moulon, Ferme du Moulon, F-91190 Gif sur Yvette, France

indices by means of estimation of the mean-square error of prediction (MSEP). We will then see how the estimated mean-square error of prediction (EMSEP) can be a robust guide in choice of a prediction strategy in various simulated situations. Finally, EMSEP will be applied to the choice of predictive traits to be included in an indirect multitrait selection index to be used at one stage of a classical soybean breeding scheme.

Evaluation of prediction quality

Interest in evaluating prediction quality

Various sources of information are available to predict the value of one single trait, and a selection index can be constructed using all of them. Often the trait of interest is observed, but associated traits observed on the same individual could be included with it in a selection index to better predict it. The same trait observed on other individuals, either genetically correlated or environmentally correlated, could bring additional information, Sometimes the phenotypic observation of the trait to be predicted is unavailable, as in the situation described in the following section.

Many prediction strategies can be used to construct an index. The observed variation could be modelled in different ways, e.g., fixed or random effects, statistical or genetical decomposition. In the same way, variance components could be estimated at different levels. For example, when studying progeny of two crosses or two populations, separate variance matrices could be estimated or they may be pooled.

Because of the wealth of potential information that could be exploited by different techniques, choice of which information to use should be made based on selection and prediction goals as well as on experimental or economical constraints. The most familiar objective is to maximize the expectation of the genetic values of retained individuals classed according to their index value. Besides this selection objective, prediction accuracy is generally of interest: the index value, $\hat{G}_i$, should be as close as possible to the real genetic value, G_i.

Classical evaluation tools and their limitations

The choice of a strategy is achieved by choosing the best index, $\hat{G}_i$, for selection (maximizing the expected genetic gain) and for prediction [minimizing the mean-square error of prediction $E(G_i - \hat{G}_i)^2$: MSEP]. Henderson (1963), Gallais (1973), and Wricke and Weber (1986) suggested the use of the prediction coefficient $R^2_{g\hat{g}}$, which proceeds from MSEP. $R^2_{g\hat{g}}$ is the linear correlation coefficient between G_i and $\hat{G}_i$, and measures the index prediction quality. Classical formulas can be used to calculate these values, which are not indepent from one another. Under known distributions, Goffinet (1983) has shown that the best index for prediction maximizes genetic progress. With such tools it is then possible to choose a genetically and economically suitable strategy.

Statistical methods allow one to make the most accurate predictions in almost all situations, using all kinds of information. Many authors (e.g., Gjedrem 1967; Gallais 1973) have shown that the more information you use, the better the prediction you get in terms of expected genetic gain. But one can rightly wonder whether multiplying predictive variables, or taking account of weakly correlated traits or individuals would not decrease the prediction accuracy. Moreover, prediction strategies require some strong assumptions, such as multivariate normality and known variance components. Discrepancies in these hypotheses are currently observed; for example, variance components are usually not known but only estimated. Harris (1964) and Sales and Hill (1976) studied the effect of errors in parameter estimates on efficiency of selection indices. Such errors lead to an over-prediction of expected genetic gain and to a loss of efficiency of the index. Sales and Hill (1976) considered the analogy with the multiple regression problem and searched for a rule on how to make a decision regarding the amount of information to be included in a selection index, i.e., a tool for index comparison. But they maintain the assumption of multinormality of the genetic values. The validity of the results is very sensitive to this assumption, especially when selection has occurred. It is therefore interesting to look for a robust criterion.

Estimated mean-square error of prediction: EMSEP

As the choice of the best selection strategy is of general interest, we have looked for a robust way to compare two indices without making any strong assumptions. We propose an unbiased estimation of MSEP that we will call EMSEP. We shall study two different situations.

Situation 1. In this case we want to select individuals i for their value G_i^1 with the observation set Y_i: $\{Y_i^1, \ldots, Y_i^k, \ldots, Y_i^K\}$ assuming the simple model:

$$Y_i = G_i + E_i$$

with: G_i: $\{G_i^1, \ldots, G_i^k, \ldots, G_i^K\}$ and E_i: $\{E_i^1, \ldots, E_i^k, \ldots, E_i^K\}$. E_i is a random array whose probability distribution is not dependent on the individual i, with null expectation, and a variance-covariance matrix Σ_E with generic term $\{\gamma_E^{kk'}\}$. This matrix is assumed to be known or estimated by an unbiased estimator, $\hat{\Sigma}_E$, with generic term $\{\hat{\gamma}_E^{kk'}\}$ obtained, e.g., by means of replications or check plots. E_i effects are assumed to be independent.

Note: the following development is easily generalizable when the selection goal is a linear combination of G_i^k.

We look for predictors $\hat{G}_i^1$ of G_i^1 having the form:

$$\hat{G}_i^1 = \alpha_i Y_i + f\{y^{[-i]}\}$$

where: α_i: $\{\alpha_i^1, \ldots, \alpha_i^k, \ldots, \alpha_i^K\}$, $y^{[-i]}$ represents all the information except Y_i, and f is any function.

This very general form contains classical BLUP predictors (Henderson 1973), where α_i^k are functions of parameters that characterize populations subject to selection. If α_i^k are fixed or independent from Y_i, we show in Appendix 1 that $EMSEP_i$ as defined below, is an unbiased estimator of $E(\hat{G}_i^1 - G_i^1)^2$:

$$EMSEP_i = (Y_i^1 - \hat{G}_i^1)^2 - \hat{\gamma}_E^{11} + 2\sum_{k=1}^{K} \alpha_i^k \hat{\gamma}_E^{1k} .$$

We will test the overall quality of a selection index by computing:

$$EMSEP = \sum_{i=1}^{N} EMSEP_i/N .$$

In actual practice, α_i^k are frequently functions of the whole set of observations, containing Y_i (variance components). It is then possible to get a good MSEP estimator, using cross-validation, i.e., computing each α_i on the sample set $y^{[-i]}$.

Situation 2: We get a sample y of individuals $i = 1, \ldots, N$ on which the array Y_i has been measured according to the model described in situation 1. In this case the term "sample" means that G_i are independent random variables with the same probability distribution. This was not required in situation 1. Moreover, E_i^K is supposed to be independent from other random variables, such as when the K variable is measured in other places, other years or, more generally speaking, under other experimental conditions.

Let us note:

$$Y_i^{[-K]}: \{Y_i^1, \ldots, Y_i^{K-1}\} .$$

We will now consider that the first $K-1$ variables are measured on a new individual, $N+1$, from the same population, that is, $Y_{N+1}^{[-K]}$. The joint probability distribution of $(Y_{N+1}^{[-K]}, G_{N+1}^K)$ is assumed to be identical to the joint probability distribution of $(Y_i^{[-K]}, G_i^K)$ for $i = 1, \ldots, N$. We are then looking for the prediction accuracy of G_{N+1}^K by $\hat{G}_{N+1}^K$, when this predictor has the form:

$$\hat{G}_{N+1}^K = \alpha\, Y_{N+1}^{[-K]} + \mu .$$

As the $N+1$ individuals come from the same population, the result of Appendix 1 indicates that each quantity, $(Y_i^K - \hat{G}_i^K)^2 - \hat{\gamma}_E^{KK}$, is an unbiased estimator of $E(G_{N+1}^K - \hat{G}_{N+1}^K)^2$, when α and μ are independent from the y sample. It becomes then natural to use the estimator:

$$EMSEP_{N+1} = \frac{1}{N}\sum_{i=1}^{N} \{(Y_i^K - \hat{G}_i^K)^2 - \hat{\gamma}_E^{KK}\} .$$

When α and μ are obtained from the same observation set, it is possible, as for situation 1, to proceed by cross-validation, i.e., computing α and μ on the sample $y^{[-i]}$ and using these estimates to calculate $\hat{G}_i^K$. But this procedure can be very expensive, so we propose to use the estimator, $EMSEP_{N+1}^*$, which corrects partially for the fact that α and μ are estimated from the same sample:

$$EMSEP_{N+1}^* = EMSEP_{N+1} + 2K\,\hat{\gamma}_E^{KK}/N .$$

Appendix 2 shows that this estimator fits well for situations where $\hat{G}^K$ is the BLUP in which parameter values are replaced by classical estimators from well-balanced situations. It is interesting to note the connection with the multiple linear regression problem. For this problem, one can use the C_p derived by Mallows (1973), to choose the number of variables to be used in the regression function. It is easily seen that the formula for C_p looks like the formula for $EMSEP_{N+1}^*$.

Use of EMSEP: an example on simulated data

EMSEP as an evaluation tool. Plantevin-Bouchez (1988) first evaluated EMSEP on simulated data, when all assumptions required by the prediction strategy were true, in a case similar to situation 1. Under these conditions, the usual tools such as MSEP and expected gain were unbiased, and necessarily more efficient than a robust estimator such as EMSEP. However, Plantevin-Bouchez (1988) verified in various genetic situations that EMSEP is unbiased in spite of a rather large sampling variance.

A reliable way to choose the best index under no strong assumptions can be achieved by minimizing EMSEP. A few arguments apply to that technique. It has been proven (Goffinet 1983) that the best prediction for MSEP is also the one ensuring the most genetic gain. This property is verified only if it is possible to attain optimality, i.e., to use the expectations of the genetic value conditional to the observations. But, if one index is better than another for the prediction, it can be worse for the selection. In particular, one should not conclude that the selection index I is poor from the fact that the prediction accuracy $EMSEP(I)$ is poor. Because to select using I or αI is equivalent (for $\alpha > 0$), one should consider I as poor only if $inf_\alpha\, EMSEP(\alpha I)$ is poor. Estimating parameters on the observation set used for constructing the index is similar to optimizing α. In such a case, minimizing EMSEP should be a reliable way of pointing out the best index for prediction and selection.

In the following we will limit discussion to applications described by situation 2. The aim is to predict individual values for one variable by means of different observed variables, when parameters have been estimated from individuals of the same population.

EMSEP for indirect multitrait selection index obtained on simulated data. A simulation study has been chosen as the

Table 1. Genetic parameter for the simulations, with $\gamma_G^{KK}=1.0$ and $\gamma_E^{kk'}=0.0$ when $k\neq k'$

No. of variables	$K=5$				$K=9$							
Predictive variables	1	2	3	4	1	2	3	4	5	6	7	8
Genotypic covariances with variable K	0.6	0.3	0.2	0.1	0.2	0.5	0.3	0.6	0.3	0.2	0.3	0.4
Genotypic variance-covariance matrix: $\Sigma_G\ (k-1, k-1)$	1.0	0.1	0.2	0.1	1.0	0.1	0.2	0.3	0.4	0.2	0.3	0.4
						1.0	0.1	0.2	0.3	0.1	0.2	0.3
		1.0	0.4	0.5			1.0	0.2	0.3	0.2	0.3	0.2
								1.0	0.2	0.4	0.2	0.3
			1.0	0.6					1.0	0.2	0.3	0.5
										1.0	0.1	0.3
				1.0							1.0	0.3
												1.0

Table 2. Mean number of predictive variables in the best index, MSEP, and observed genetic gain for different simulated observation sets (N is the number of individuals, γ_E^{kk} is environmental variance, ** is significant at 1% probability level) for $K=5$ variables

N	γ_E^{kk}	Mean no. of predictive variables	MSEP			Genetic gain		
			Best index	Full index	Difference	Best index	Full index	Difference
20	0.5	2.0	1.477	1.488	NS	2.328	2.416	NS
	1.0	1.6	2.123	2.184	**	1.691	1.710	NS
	1.5	1.4	2.718	2.847	**	1.335	1.272	NS
60	0.5	2.4	1.285	1.286	NS	9.878	9.926	NS
	1.0	2.0	1.883	1.887	NS	7.817	7.897	NS
	1.5	1.7	2.450	2.460	**	6.373	6.422	NS

simplest way to validate the efficiency of EMSEP for choosing, from among many combinations of predictive variables, the best index for the variable to be predicted. In order to decrease computation time, data has been simulated according to a simple balanced situation.

From a multinormal probability distribution, we draw the observations array Y_i: $\{Y_i^1, \ldots, Y_i^k, \ldots, Y_i^K\}$ for $i=1, \ldots, N$ according to the simple model:

$$Y_i^k = G_i^k + E_i^k$$

for various genetical and environmental situations (Table 1). In this model, the E_i^K are independent from other environmental values E_i^k. The phenotypic covariance is then equal to the genotypic one. It is then possible, with the phenotypical variance-covariance matrix $\hat{\Sigma}_P$ estimated from Y_i, to predict the $\hat{G}_i^K$ by means of the observed values $Y_i^{[-K]}$ for all possible combinations of 1 to $K-1$ predictive variables. The best index is selected as the combination that minimizes EMSEP* computed on the N $\hat{G}_i^K$ values.

In a second step, a new sample Y_i is drawn for $i=N+1, \ldots, 2N$. $\hat{G}_i^K$ are then estimated by two different indices: (*i*) the best index calculated in step 1, and (*ii*) the complete index using the $K-1$ predictive variables with parameters estimated from step 1. MSEP and genetic gain (for a selection rate of 30%) are computed for both indices. This second step is repeated 20 times, and the comparison of indices is based on mean values of MSEP and genetic gain. Tables 2 and 3 show mean results obtained on 300 simulations for different values of N, K and Σ_E.

As one would expect, the best index according to EMSEP* ensures better prediction accuracy than the index using all the variables. This decrease of MSEP is observed in all simulated cases, and the MSEP difference is significant in most cases, especially when using $K=9$ variables.

As the environmental part of total variance increases, the accuracy of estimating Σ_G decreases, bringing about a decrease in prediction quality for $\hat{G}_i^K$. The MSEP difference between the indices grows and, consequently, increases the interest in an index using fewer predictive variables.

When the individual number N is changed from 20 to 60, parameter estimation is more precise, consequently decreasing MSEP and minimizing the difference in prediction accuracy between the two types of indices.

On the other hand, the index choosen by means of EMSEP* rarely leads to more genetic gain and, in a few cases ($K=9$, $N=60$), to significantly smaller gain than the one obtained by means of the complete index; nevertheless, this decrease in genetic gain remains small. It is important to mention that these simulations were for a balanced situation that is quite favorable to classical cal-

Table 3. Mean number of predictive variables in the best index, MSEP, and observed genetic gain for different simulated observation sets (N is number of individuals, γ_E^{kk} is environmental variance, ** is significant at 1% probability level) for $K=9$ variables

N	γ_E^{kk}	Mean no. of predicitve variables	MSEP			Genetic gain		
			Best index	Full index	Difference	Best index	Full index	Difference
20	0.5	3.5	1.750	1.889	**	2.498	2.469	NS
	1.0	2.7	2.538	2.891	**	1.755	1.741	NS
	1.5	2.3	3.300	3.849	**	1.315	1.367	NS
60	0.5	4.3	1.241	1.240	NS	11.151	11.320	**
	1.0	3.5	1.888	1.899	**	8.927	9.106	**
	1.5	3.0	2.500	2.528	**	7.186	7.463	**

culation of MSEP. In such cases, the EMSEP correction plays a very important part, and the observed decrease in genetic gain let us think that this correction may be improved. It would be interesting to be able to understand why the index choice based on EMEP* does not lead to the intended increase in genetic gain.

However, the main interest for finding the best index becomes obvious. In every case an index including fewer predictive variables than the complete one is pointed out by EMSEP*. Such an economy of variables always leads to better prediction quality, and rarely to a significant decrease in genetic gain. As potential estimation problems for parameters become more important (high γ_e^{kk} or small N), EMSEP* leads to an even smaller number of predictive variables. On the other hand, as estimations are more and more likely to be close to real parameter values, it becomes more and more interesting to use all the available information for prediction of individual genetic values. Nevertheless, regardless of the situation, the EMSEP* diagnosis has allowed us to decrease significantly the number of variables going, in one case ($K=5$, Table 2), from four to one or two variables, in the other case ($K=9$, Table 3), from eight to about three variables.

Choice of an indirect multitrait selection index for soybean improvement, by means of EMSEP

Selection among F_5 single plants derived by SSD

Single seed descent (SSD) is a strain fixing method that is commonly used for soybean improvement (Fehr 1978). Homozygosity is quickly increased without any selection up to the F_5 or F_6 generation, maintaining in this way all of the genetic variation. However, going from this fixation stage to the selection stage remains a real problem for soybean breeders. How to make a first choice among the large number of strains on the basis of observation of one single F_5 or F_6 plant by strain? The two important variables to be predicted are yield and precocity of fixed strains but, to simplify, we consider only the prediction of precocity. In order to do that, it is possible to look for good correlations between the variable to be predicted and one variable or a combination of variables that could be observed on single F_5 plants, constructing in this way a selection index allowing an early efficient choice.

Progenies from three crosses between soybean varieties of indeterminate type have been brought to F_5 by SSD. Nine variables have been observed on 100 F plants from each cross: production variables; total number of pods (NPT), number of pods per ramification (NPR), number of seeds (NS), total seed yield (YD), total dry matter (DM), and developmental variables; number of ramifications (RAM), number of nodes (NO), plant height (HT), maturity date (R8). Strains from 30 F_5 plants selected among the 100 in each cross were observed for precocity in agronomic trials with replications of F_7 and F_8, in 2 successive years.

Estimation and prediction methods

The observations Y_{ij}^k of k variables ($k=1,\ldots,9$) on F_5 plants follow a multinormal probability distribution, and are described by the following mixed model, for $i=1,\ldots,100$ individuals and $j=1,\ldots 3$ crosses:

$$Y_{ij}^k = \mu_i^k + G_{ij}^k + E_{ij}^k .$$

Observations Y_{ijl}^{hK} from F_7 ($h=7$) or F_8 ($h=8$) strains, for the variable K to be predicted (precocity), made in an incomplete block design, are described by the following model:

$$Y_{ijl}^{hK} = \mu_j^{hK} + B_l^{hK} + G_{ij}^{hK} + E_{ijl}^{hK}$$

where the μ_j^k are the fixed effect related to the j cross, the B_l^{hK} are the random effect related to the l block (trial), the G_{ij}^k are the genetic random effects related to the i individual from the j cross, and the E are the residual effects.

We assume that $E_{ij}=(E_{ij}^1,\ldots,E_{ij}^k)$, $G_{ij}^h=(G_{ij}^1,\ldots,G_{ij}^{K_1}, G_{ij}^{hK})$, E_{ijl}^{hK}, and B_l^{hK} are, respectively, multinormally distributed and independent of all the other random variables.

Note that genetic variances among the strains of the three different croses are asumed to be homogeneous, and a single variance-covariance matrix has been estimated for all strains.

The predictions of F_7 using F_5, and of F_8 using F_5 have been considered independently. In each case, parameters have been estimated by the maximum likelihood method, using the iterative procedure proposed by Rao and Kleffe (1980) on the whole set of data, respectively, F_5 and F_7, or F_5 and F_8. The maximum likelihood estimators based on all the information are, in fact, the only ones that lead to good predictors when selection has occurred in the observed population (Thompson 1979). We need to assume that selection has occurred only in the F_5 generation.

For $h=7$ (F_7), and $h=8$ (F_8), we want to predict $\mu_j^{hK}+G_{ij}^{hK}$ in each cross j, for $i=1,\ldots,100$ strains. The indices that have been used are based on the Best Linear Unbiased Predictor (BLUP) (Henderson 1963), for all possible combinations of 1–9 predictive variables of F_5, $Y^{[-K]}$, constructed with estimated parameters.

Index evaluation

As parameters are not known but estimated, classical calculation formulas for prediction accuracy, MSEP, and genetic gain have no reason for being unbiased. In such cases, EMSEP* is the only unbiased comparison criterion. Indices have, therefore, been compared by means of their EMSEP* values to determine the best index for each number of variables. This has been achieved by looking for the best univariable index, combining then the variable that leads to the best bivariable index, and so on, up to nine variables. It then becomes possible to choose the best combination of predictive variables.

It should be noted that the assumptions of normality, used to build the estimators, are not needed for the unbiasedness of the EMSEP* value.

Results

Figures 1 and 2 give MSEP and EMSEP* values for combinations of 1–9 observed F_5 variables, for the prediction of F_7 and F_8 precocity. The order of introduction of F_5 variables in best indices is similar for both experimental years, and brings out the importance of two predictive variables: plant height (HT) and maturity date (R8). However, as shown in Table 4 it is not the same index that minimizes EMSEP* in both cases: for F_7 the best prediction is based solely on HT, and for F_8 it is a six-variable index that leads to the best prediction accuracy. F_7 and F_8 were grown under very contrasting climatic conditions, which could explain the difference in best indices. A bivariate index combining HT and R8 should be rather easy to obtain from a practical point of view, and should provide good prediction quality in all years.

Interest for EMSEP

In all studied cases, EMSEP enables us to select an index with a reduced number of predictive variables (Figs. 1 and 2). We notice from these figures that values for EMSEP and MSEP are often quite different, and that MSEP always leads to use of all the available information.

EMSEP appears to be a robust and original tool for choosing variables to be used in an index. It is obvious that adding predictive variables is not always desirable,

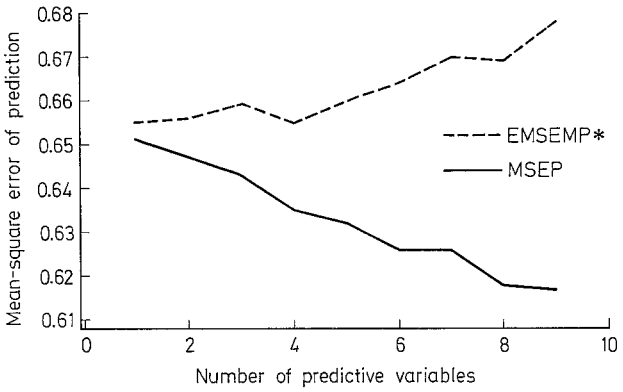

Fig. 1. *F_7 precocity prediction:* MSEP and EMSEP* values for best indices (order of introduction of predictive variables: 1-HT, 2-R8, 3-NPT, 4-NO, 5-NS, 6-RAM, 7-YD, 8-DM, 9-NPR)

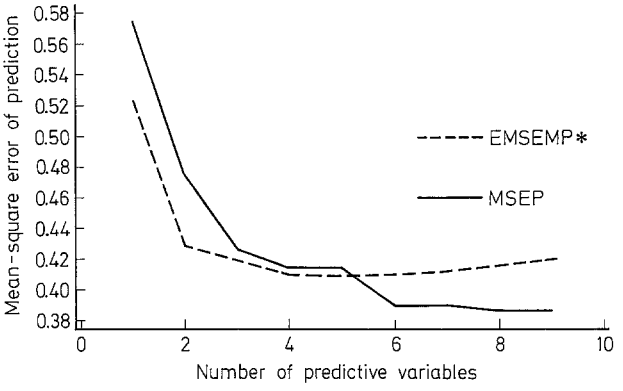

Fig. 2. *F_8 precocity prediction:* MSEP and EMSEP* values for best indices (order of introduction of predictive variables: 1-R8, 2-HT, 3-NS, 4-YD, 5-NPT, 6-DM, 7-RAM, 8-NO, 9-NPR)

Table 4. Precocity prediction: EMSEP* for some interesting combinations of F_5 variables

Index	Variable no.	EMSEP*	
		F_7	F_8
Complete index	9	0.677	0.420
Best F_7 index HT	1	0.655	0.557
Best F_8 index R8-HT-NS-YD-NPT-DM	6	0.660	0.407
Intermediate index HT-R8	2	0.656	0.436

as it may lead to problems in estimation of parameters and to a decrease in prediction accuracy.

However, these results apply only to the studied samples, and do not resolve completely the problem of year effect. The use of indicated indices for another sample, progenies from other crosses, e.g., is not obvious. We may hope to apply them to populations in which parents are not too different from those studied here. In any case, this should be done with care, using new estimates of variance parameters from F_5.

First and foremost, our aim is to propose EMSEP as a tool for choosing among predictive variables as has been required by Sales and Hill (1976), this tool being robust to discrepancies in assumption of normality and parameter estimates.

Appendix 1

The following equalities are easy to verify:

$$E(\hat{G}_i^1 - Y_i^1)^2 = E((\hat{G}_i^1 - G_i^1) - (Y_i^1 - G_i^1))^2$$
$$= E(\hat{G}_i^1 - G_i^1)^2 - 2E(\hat{G}_i^1 - G_i^1)(Y_i^1 - G_i^1) + E(Y_i^1 - G_i^1)^2$$
$$E(\hat{G}_i^1 - G_i^1)^2 = E(\hat{G}_i^1 - Y_i^1)^2 - \gamma_E^{11} + 2\sum_{k=1}^{K} \alpha_i^k \gamma_E^{1k} .$$

If $(\hat{G}_i^1 - Y_i^1)^2$ is an unbiased estimator of $E(\hat{G}_i^1 - Y_i^1)^2$, then it is possible to get an unbiased estimator of $E(\hat{G}_i^1 - G_i^1)^2$:

$$EMSEP_i = (\hat{G}_i^1 - Y_i^1)^2 - \hat{\gamma}_E^{11} + 2\sum_{k=1}^{K} \alpha_i^k \hat{\gamma}_E^{1k} .$$

Notice that the validity of this expression does not depend on the probability distribution of G_i^1, as all equalities are valid conditional on the G_i^1 value, though there is no need for the hypothesis of random G_i^1, even if it had played some part in the construction of the prediction, $\hat{G}_i^1$.

Appendix 2

Let us suppose there is a true model:

$$Y_i^K = a + b\, Y_i^{[-K]} + \varepsilon_i$$

where $\mathrm{Var}(\varepsilon_i) = \gamma_\varepsilon$ does not depend on i.

Let $\hat{a}$ and $\hat{b}$ be least-square estimates of a and b. The μ and α coefficients for index $\hat{G}^K$ are, respectively, $\mu = \hat{a}$ and $\alpha = \hat{b}$ when the index is the best linear unbiased predictor, and where variance-covariance parameters values are replaced by their estimates from the sample y. In such a situation, we are looking for the estimation of R^K:

$$R^K = E(G_{N+1}^K - \hat{G}_{N+1}^K)^2$$
$$= E(G_{N+1}^K - (\hat{a} + \hat{b}\, Y_{N+1}^{[-K]}))^2$$
$$= E(G_{N+1}^K - (a + b\, Y_{N+1}^{[-K]}))^2 + E(\hat{a} + \hat{b}\, Y_{N+1}^{[-K]} - (a + b\, Y_{N+1}^{[-K]}))^2$$
$$= R_1^K + R_2^K .$$

It is evident that $R_1^K = \gamma_\varepsilon - \gamma_E$.

For the second part, R_2^K, let us first make the approximation that the probability distribution of $Y_{N+1}^{[-K]}$ is the discrete probability distribution taking each $Y_i^{[-K]}$ value with a probability $1/N$. Using the algebraic form of least-square estimates, $\hat{a}$ and $\hat{b}$, it is easily demonstrated that $R_2^K = \gamma_\varepsilon K/N$. The effect of this approximation can be measured by calculating the exact value of R_2^K when $K = 2$ and when $Y_i^{[-K]}$ is assumed to be normally distributed. This leads to an independent χ^2 ratio, and gives in the end $R_2^K = (\gamma_\varepsilon K/N)(N/(N-2))$. The expression $R_2^K = \gamma_\varepsilon K/N$ seems then to be good enough for usual sample sizes in plant breeding.

On the other hand, let us look for the expectancy of:

$$\hat{R}^K = (1/N)\,\Sigma_i \{(Y_i^K - (\hat{a} + \hat{b}\, Y_{N+1}^{[-K]}))^2 - \hat{\gamma}_E^K\} .$$

Linear model theory shows that:

$$E[\hat{R}^K] = \gamma_\varepsilon \frac{N-K}{N} - \gamma_E .$$

So $\hat{R}^K + 2K\,\gamma_\varepsilon/N$ is an unbiased estimator for R^K. The expression for $EMSEP_{N+1}^*$ is obtained by replacing γ_ε by $\hat{\gamma}_E^{KK}$, which is a biased estimator for γ_ε, but has been shown with simulation to be more satisfactory than the use of an unbiased estimator with a big sampling variance.

References

Fehr WR (1978) Breeding. In: Norman AG (ed) Soybean: physiology, agronomy and utilization. Academic Press, London New York, pp 119–155

Gallais A (1973) Sélection pour plusieurs caractères: synthèse critique et généralisation. Ann Amel Plant 23:183–208

Gianola D (1986) On selection criteria and estimation of parameters when the variance is heterogeneous. Theor Appl Genet 72:671–677

Gjedrem T (1967) Selection indexes compared with simple trait selection. I. The efficiency of including correlated traits. Acta Agric Scand 19:263–268

Goffinet B (1983) Selection on selected records. Genet Sel Evol 15:91–97

Goffinet B (1987) Alternative conditions for ignoring the process that causes missing data. Biometrika 74:437–439

Harris DL (1964) Expected and predicted progress from index selection involving estimates of population parameters. Biometrics 20:46–72

Henderson CR (1963) Selection index and expected genetic advance. In: Hanson WD, Robinson HF (eds) Statistical genetics and plant breeding. NAS-NRC Publication No. 982, Washington/DC, pp 141–163

Henderson CR (1973) Sire evaluation and genetic trends. In: Proc Anim Breed Genet Symp in Honor of JL Lush. Am Soc Anim Sci, Am Dairy Sci Assn, Champaign/IL, pp 10–41

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Henderson CR (1986) Recent developments in variance and covariance estimation. J Anim Sci 63:208–216

Mallows CL (1973) Some comments on Cp. Technometrics 15:661–675

Plantevin-Bouchez A (1988) Contribution à la construction et à la validation des index de sélection, Application à la sélection du soja. PhD Thesis, Institut National Polytechnique de Toulouse, France

Quaas RL, Pollak EJ (1980) Mixed model methodology for farm and ranch beef cattle testing programs. J Anim Sci 51: 1277–1287

Rao CR, Kleffe J (1980) Estimation of variance components. In: Krishnaiah PR (ed) Handbook of statistics 1, analysis of variance. NHPC, Amsterdam New York Oxford, pp 1–40

Sales J, Hill WG (1976) Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Anim Prod 23:1–14

Thompson R (1979) Sire evaluation. Biometrics 35:339–353

Wricke G, Weber WE (1986) Quantitative genetics and selection in plant breeding. de Gruyter, Berlin New York