Genet Sel Evol (1995) 27, 551-565

Original article

Approximating selection differentials and variances for correlated selection indices

F Phocas, JJ Colleau

Institut national de la recherche agronomique, station de génétique quantitative et appliquée, 78352 Jouy-en-Josas cedex, France

(Received 1st June 1994; accepted 1st August 1995)

Summary – Empirical formulae were derived to approximate selection differentials and variances of the selected estimated breeding values when the estimated breeding values of the candidates for directional selection are multinormally distributed and correlated in any manner. These formulae extended the well-known exact basic form for the equicorrelated case, taking into account selection pressure, average pairwise correlation coefficient and average standard deviation of pairwise correlation per observation, through polynomials fitted to simulated data. Simulations were carried out for different correlation structures (1, 2 or 3 different intra-class correlations per family, ranging from 0.3 to 0.99), for different numbers of independent families (1, 2, 5 or 10), for constant or variable family size and for selection pressures ranging from 0.5 to 50%. On average, 90% of the bias occurring when ignoring correlations between observations was removed by our prediction formula of selection differential or variance of selected observations. Comparisons with other correction methods, which assume special correlation structures, were also carried out.

selection differential / correlated indices / finite population

Résumé – Approximations empiriques des différentielles de sélection et des variances pour des indices de sélection corrélés. *On propose des formules de calcul approché des différentielles de sélection et des variances d'index de sélection après sélection directionnelle quand les candidats à la sélection ont des index distribués normalement et corrélés de manière quelconque. Ces formules ont pour base celles établies en cas d'équicorrélation entre observations et font intervenir des polynômes des variables suivantes : taux de sélection, coefficient de corrélation moyen et écart type moyen de ce coefficient par observation. Les coefficients des polynômes sont calculés après ajustement à des données simulées. Les situations simulées font varier la structure des corrélations (1, 2 ou 3 coefficients de corrélation intra-classe, de valeurs 0,3 à 10,99), le nombre de familles (1, 2, 5 ou 10), la taille de famille (constante ou non) et le taux de sélection (de 0,5 à 50%). En moyenne, 90% du biais introduit en ignorant les corrélations entre observations est corrigé par nos*

formules de prédiction des différentielles de sélection et des variances des observations sélectionnées. Des comparaisons sont effectuées avec d'autres méthodes de correction proposées pour des structures de corrélation particulières.

différentielle de sélection / indices corrélés / population finie

INTRODUCTION

The relative efficiencies of alternative breeding schemes can be assessed through deterministic predictions. Both selection and limited size of breeding populations lead to complex consequences for genetic gains, so that unbiased predictions are difficult to obtain (see review by Verrier *et al*, 1991, for example). An important consequence is that estimated breeding values (EBVs) of candidates are correlated (through genetic relationships and for statistical reasons, because EBVs are obtained from the same set of observations). However, a very common assumption is that candidates correspond to independent observations from an infinite population. Consequently, genetic gains are overestimated because selection differentials and variances of EBVs between selected candidates are overestimated. The amount of bias can differ according to breeding scheme and correctness of comparisons between schemes can be impaired.

Burrows (1972) provided an accurate, easy to implement, approximation of selection differentials when independent candidates are drawn from a finite population. When the number of observations is larger than 5, it leads to errors which are always smaller than 2%, and usually smaller than 1%. Conversely, no exact method has been found to take into account any correlated structure among normally distributed observations. Owen and Steck (1962) gave the exact solution for equicorrelated multivariate normal distribution. If we define uniform families as families of identical size and identical within-family correlation structure, Hill (1976) and Rawlings (1976) provided the exact solution for the case of uniform independent families of within-family equicorrelated observations. Since this solution uses multiple numerical integration, they proposed *ad hoc* approximations which were relatively poor for high intra-class correlations (over 0.6) and severe selection pressures (below 10%). Rawlings' empirical formula was based on Owen and Steck's result for the equicorrelated case. Perez-Enciso and Toro (1991) proposed a method to account for any variance–covariance structure among indices. For equal variances, their method corresponded to Rawling's approximation. Meuwissen (1991) improved Rawlings' approximation for the case of several uniform families and found an extension for uniform full-sib families nested within uniform half-sib families. His correction was very accurate for the breeding schemes examined, *ie* assuming a hierarchical mating design. However, it cannot be generalized to any correlation structure.

The purpose of this paper is to provide approximation formulae for both the selection differential and the variance of EBVs of selected candidates, assuming no specific correlation structure but assuming that variances of EBVs are constant

across candidates. Keeping Meuwissen's basic idea, these formulae are derived by fitting an extended Owen and Steck's formulae to simulated data.

PREDICTION FORMULAE

General form

Selection differential

Rawlings' (1976) formula consists of using Owen and Steck's (1962) exact formula for selection differential when population is split into independent and uniform families:

$$I_R = I_0(1 - r)^{0.5}$$

I_0 is the standardized selection differential for finite independent observations and depends on n (number of candidates), p (selection rate), I_∞ (selection differential for infinite independent observations) through Burrows' approximation:

$$I_0 = I_\infty - \frac{1 - p}{2(n + 1)pI_\infty}$$

r is the average pairwise correlation coefficient and is equal to

$$\rho\frac{s - 1}{fs - 1}$$

for f families of size s with within-family correlation coefficient ρ.

We suggest here a generalization of Rawlings' formula for any correlation structure, taking into account the following parameters:

1) the selection pressure ($p \leqslant 0.5$);
2) the average pairwise correlation coefficient:

$$r = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\rho_{ij}$$

where n is the number of candidates and ρ_{ij} is the correlation between EBVs of candidates i and j;

3) the average standard deviation of the pairwise correlation coefficients involving a given candidate

$$\sigma_r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-2}\sum_{j\neq i}\left(\rho_{ij} - \frac{1}{n-1}\sum_{j\neq i}\rho_{ij}\right)^2\right)^{0.5}$$

When variances of EBVs are standardized to 1, the analytical expression of the approximation proposed is:

$$I_p = I_0(1 - r)^{0.5+\sigma_r P(p,r,\sigma_r)} \qquad [1]$$

where P stands for a polynomial of variates p, r, and σ_r. In the equicorrelation situation ($\sigma_r = 0$), Owen and Steck's exact results still hold with such an approximation.

Rawlings (1976) compared his approximate correction with exact results obtained from numerical integration and found that the discrepancies between them increased when correlations increased. This justified a further correction term including r. Introduction of parameters σ_r and p was basically justified by the fact that Rawlings' approximation is less and less accurate when the variability of ρ_{ij} increases and/or p decreases. The polynomial form of the approximation was considered to be the simplest to implement when no analytical underlying theory is referred to.

Variance of selected EBVs

Owen and Steck's (1962) exact result for the equicorrelated case is $V_c = (1 - r)V_0$ where V_0 is the variance of selected independent observations. Burrows (1972) showed that population size hardly affects this last variance. Therefore, V_0 is calculated as for an infinite population.

$$V_0 = 1 - I_\infty(I_\infty - x_\infty)$$

where x_∞ is the selection threshold in an infinite population.

The analytical expression of the approximation is:

$$V_Q = (1 - r)V_0(1 + Q(p, r, \sigma_r)) \qquad [2]$$

where Q is a polynomial of variates p, r, σ_r cancelling out when $\sigma_r = 0$.

Data examination showed that the first part of the approximation, $V_c = (1 - r)V_0$ accounted for the major part of variance reduction induced by a correlated structure. The second part of the approximation was introduced as a multiplier factor because observation on calibration data sets showed that this method provided positive approximations for variances. Expressions ensuring positivity in any situation, such as $(1 - r)V_0$ exp(polynomial), were not able to provide a good fit. We will comment further on this point.

Fitting polynomial coefficients

Different structures were generated to provide variation for r and σ_r. For a given structure, 5 000 replicates were generated. Subsamples corresponding to different selection rates (p) were extracted. The basic observed values I_{obs} and V_{obs} were respectively the averaged values of selected candidates (selection differentials) and the pooled value of within-replicate variances of selected candidates.

Only p values equal to or lower than 0.50 were investigated since the following equations exist:

$$I_{1-p} = \frac{p}{1-p} I_p$$

$$V_{1-p} = \frac{1}{1-p}\left(1 - pV_p - \frac{p}{1-p} I_p^2\right)$$

Therefore, if p were greater than 0.5, the prediction should hold for $p^* = 1 - p$ and back solution for p would be given by the above formulae.

Dependent combined observed values from several combinations of data structure × selection rate were analysed to test a polynomial regression, using the SAS procedure 'General Linear Models' (SAS/STAT User's Guide, 1990).

To estimate coefficients of the polynomial P, the dependent variate y was such that:

$$I_{\text{obs}} = I_0(1-r)^{0.5+\sigma_r y}$$

which corresponded to

$$y = \frac{1}{\sigma_r}\left[\left[\ln\left(\frac{I_{\text{obs}}}{I_0}\right)/\ln(1-r)\right] - 0.5\right]$$

For the polynomial Q, dependent variate z was such that $V_{\text{obs}} = (1-r)V_0(1+z)$ which corresponded to

$$z = \frac{V_{\text{obs}}}{V_0(1-r)} - 1$$

Testing goodness of fit

Polynomials of degrees 5 and 6 were tested for P and Q, respectively. They provided better adjustments (R-square values) than polynomials of lower degrees. Fitting higher polynomials led to singularities in our data sets.

Only significant polynomial coefficients on p, r, σ_r and higher degrees of these variates were considered for use in correction formulae.

In addition to the R-square values provided by the model, relative errors incurred with different procedures were considered:

– from treating variates as independent

1) for selection differentials $U_I = 100\dfrac{I_0 - I_{\text{obs}}}{I_{\text{obs}}}$

2) for variance of selected observations $U_V = 100\dfrac{V_0 - V_{\text{obs}}}{V_{\text{obs}}}$

– from correction attempts according to different formulae

1) for selection differentials $F_I = 100\dfrac{|I_F - I_{\text{obs}}|}{I_{\text{obs}}}$

where F is a generic letter corresponding to R, M, P (Rawlings, Meuwissen and polynomial formulae, respectively)

2) for variance of selected observations $F_V = 100\dfrac{|V_F - V_{\text{obs}}|}{V_{\text{obs}}}$

with F corresponding to B (Owen and Steck, 1962) or P (polynomial formula).

Absolute values of ratios are used because correction formulae sometimes lead to overcorrection, *ie* negative values of relative errors. Rawlings' formula often corresponds to overestimation. Regression formulae such as Meuwissen's and polynomial formulae lead to overestimates in some cases or underestimates in others.

Correction inefficiency corresponds to the ratio of errors still remaining after correction, compared with errors incurred with no correction at all.

Correction inefficiencies for selection differentials correspond to ratios $F_I^* = F_I/U_I$, where F stands for alternative correction formulae. Correction inefficiencies for variances correspond to ratios $F_V^* = F_V/U_V$. When reading tables, small values are favourable when considering either errors or correction inefficiencies.

SIMULATED DATA SETS

Calibration data sets

Two sets of simulated data were generated and pooled to estimate coefficients of the polynomials involved in the previous formulae. These data sets were chosen in order to represent a large variation for the correlation structure among EBVs. For that purpose, values of intra-class correlations were arbitrarily taken without considering real breeding scheme structures. An n-candidate layout was simulated as a set of n correlated standardized normal variates, the basic normal variates representing EBVs. In such a simulation, there is no need to simulate performances leading to these EBVs.

Data set 1

In the first data set, 40 candidates for selection were simulated; 1 200 situations were examined according to the number (1, 2 or 5) of independent groups, called 'families', and the size of these groups (constant or variable). Possible contributions of families, when family size is not constant, are shown in table I. Selection pressures were 50, 40, 30, 20, 10 and 5%. Furthermore, 3 correlation structures were simulated.

Table I. Distribution of candidates according to families.

Family	*Data set 1*		*Data set 2*				
1	80%	60%	80%	60%	40%	55%	35%
2	20%	10%	20%	10%	20%	5%	15%
3		10%		10%	20%	5%	15%
4		10%		10%	10%	5%	5%
5		10%		10%	10%	5%	5%
6						5%	5%
7						5%	5%
8						5%	5%
9						5%	5%
10						5%	5%

In the first correlation structure, candidates of the same family were equicorrelated. This could correspond to full-sib of half-sib family structures. Cases with 1 family were not simulated since the exact result is known. Intra-class correlation values considered are shown in table II.

Table II. Correlation structures.

Structure	*Data set 1*	*Data set 2*
One correlation within family	0.99	0.99
	0.9	0.7
	0.7	0.3
	0.5	
	0.3	
Two correlations within family	0.99 – 0.7	0.99 – 0.7
	0.99 – 0.5	0.9 – 0.5
	0.99 – 0.3	0.5 – 0.3
	0.9 – 0.7	
	0.9 – 0.5	
	0.9 – 0.3	
	0.7 – 0.5	
	0.7 – 0.3	
	0.5 – 0.3	
Three correlations within family	0.99 – 0.9 – 0.7	0.99 – 0.9 – 0.7
	0.99 – 0.9 – 0.5	0.9 – 0.7 – 0.5
	0.99 – 0.9 – 0.3	0.7 – 0.5 – 0.3
	0.9 – 0.7 – 0.5	
	0.9 – 0.5 – 0.3	
	0.7 – 0.5 – 0.3	

In the second correlation structure, 2 different intra-class correlations were considered within each family. This corresponded to the nested full-half sib family structure analysed by Meuwissen (1991); each family of half-sibs was made up of several groups of full-sibs. The number of full-sibs was 2 in each group. The 9 considered pairs of intra-class correlation coefficients between full- and half-sibs are shown in table II.

In the third correlation structure, 3 different intra-class correlation values per family were considered because each family was split into 2 subgroups. Correlations within sub-groups were r_{11} and r_{22} respectively. Correlation between sub-groups was r_{12}. This could correspond to subgroups with different information, although strictly speaking, this would lead to heterogeneity of variance. The 6 combinations (r_{11}, r_{22}, r_{12}) considered are shown in table II. Sub-group 1 represented 25, 50 or 75% of each family.

Data set 2

In the second data set, 315 situations involving many more candidates (200) and more severe selection pressures (0.5, 1, 5%) were simulated. The number of families was 1, 2, 5 or 10. Heterogeneity for family size is shown in table I. Correlation structures varied according to the same principle as in data set 1 but values were not quite the same (see table II). Sub-group 1 represented 25 or 75% of each family.

Cross-validation data sets

The aim of these data sets is to validate the prediction formulae for correlation structures different from those used for fitting the polynomials. This is a way to test the prediction abilities and robustness of the fitted polynomial equations.

Four situations relative to breeding schemes (10 000 replicates per situation) were considered to derive different structures of correlations among indices. A BLUP animal model evaluation was used to rank animals.

Beef cattle breeding schemes (2 situations)

Correction formulae were tested on a simulated selection nucleus for beef cross-breeding on dairy cattle (Phocas *et al*, 1995, unpublished results).

Generation 1 consisted of 186 dams born from 31 unrelated sires. Generation 2 was produced by mating these dams to 3 sires (1 calf per dam). A BLUP evaluation was implemented, assuming that males or females of generation 2, females of generation 1, males of generation 1, and males of generation 0 were recorded for a single trait.

The first situation corresponded to $h^2 = 0.25$ and the second corresponded to $h^2 = 0.10$.

The efficiency of our correction formulae was tested on generation 2, when selecting replacement females ($p = 46/93$) and males ($p = 3/93$) and for both heritabilities. Candidates for selection can be unrelated, half-sibs (same sire), cousin (same maternal grand-sire) or both at the same time. For $h^2 = 0.25$, values for r and σ_r were 0.176 and 0.246. For $h^2 = 0.10$, the corresponding values were 0.223 and 0.309.

Dairy cattle breeding schemes (2 situations)

Intensive breeding schemes using embryo transfer and putting emphasis on pedigree selection are likely to induce high correlations between EBVs of candidates and to reduce effective selection differentials. These schemes are referred to as multiple ovulation and embryo transfer (MOET) schemes (Nicholas and Smith, 1983).

The efficiency of the proposed correction formulae was tested on the 192 females of generation 2, born from 4 sires and 48 dams. Each dam was mated to 2 different sires (factorial mating design). Each mating produced 2 females (and 2 males). The 48 dams were assumed to be recorded on milk yield ($h^2 = 0.25$) and to be born from 4 sires, unrelated to sires of generation 2. Female candidates of generation 2 were assumed to be evaluated according to a BLUP procedure, to produce replacement females or replacement males.

An 'adult' MOET (first situation) was mimicked assuming generation 2 was recorded (1 lactation per individual). In this situation, relevant r and σ_r were 0.160 and 0.220, respectively. If a 'juvenile' MOET (second situation) was implemented, females of generation 2 were not recorded before selection. In our layout, all the progeny (4 individuals) of the same dam had the same EBV. Therefore, selection was not carried out among 192 individual EBVs but among 48 EBVs groups; the corresponding r and σ_r were 0.137 and 0.251.

RESULTS

Fitting selection differentials

Polynomial P was estimated from the observed data on 1 515 (*ie* 1 200 + 315) basic situations. However examination of the results showed that high values of $r (r > 0.6)$ were detrimental to goodness of fit. Therefore, we restricted data adjustment to the 1 383 situations where r was smaller than 0.6.

Coefficients of the polynomial of degree 5 shown in the *Appendix* were found to be significant. Similarity of coefficients suggested some grouping and the variate transformation $d = r - \sigma_r$. Examination of the new results suggested further additional variate transformations $e = r(1 - r)$ and $b = d(1 - 4.2e)$. This led to only 5 significant regression coefficients without loss of accuracy, as compared to the first adjustment. This polynomial was:

$$P(p, \sigma_r, b) = \sigma_r^2(c_1 + c_2 p\sigma_r) + b(c_3 + c_4 p + c_5 p^2) \quad [3]$$

with $c_1 = 7.6$ (0.4)
$c_2 = -30.5$ (3.6)
$c_3 = -55.8$ (0.8)
$c_4 = 375.3$ (9.6)
$c_5 = -557.5$ (19.4)

where estimation standard errors are in parentheses. In these conditions, the R-square value for this polynomial adjustment was found to be 0.85.

Table III shows that the average relative error (P_I) was only 2.5% compared with 26.4% when no correction is used and with 7.1% when Rawling's correction is made. In 96% of cases, relative errors were smaller than 10%, whereas this occurred in only 20% of cases when no correction was used and 79.5% of cases when Rawling's correction was implemented. The average correction inefficiency rate of the polynomial adjustment (P_I^*) was 10%, which meant that 90% of the bias occurring with no correction for correlated EBVs was removed. Only 77% of this bias was removed by Rawlings' formula.

Table III. Comparison of Rawlings' and polynomial formulae (1 383 situations with average pairwise coefficient smaller than 0.6).

	U_I	R_I	P_I	R_I^*	P_I^*
Mean	26.4	7.1	2.5	0.23	0.10
Standard deviation	21.9	11.2	3.2	0.18	0.09
Maximum value	237.6	138.3	28.3	0.85	0.46
% less than 10%	20.0	79.5	96.0	28.1	60.5
% less than 20%	47.4	91.5	99.8	54.2	89.0

U_I = relative error on selection differential with no correction; R_I, P_I = relative error when correcting by Rawlings' and polynomial formulae, respectively; R_I^*, P_I^* = correction inefficiency rate by Rawlings' and polynomial formulae, respectively.

Table IV shows, however, that quality of adjustment was still dependent on p values. For small p ($p < 5\%$, *ie* 478 situations out of 1 383), the average relative error was 3.8%. This value compared favourably with corresponding figures for no correction (31.1%) or Rawlings' correction (12.3%).

Table IV. Comparison of Rawlings' and polynomial formulae according to selection rate p.

	U_I	R_I	P_I	U_I^*	P_I^*
$p \geqslant 10\%$; 905 situations					
Mean	23.9	4.3	1.9	0.16	0.09
Standard deviation	17.6	6.1	2.3	0.12	0.08
Maximum value	147.5	78.0	15.6	0.65	0.46
$p \leqslant 5\%$; 478 situations					
Mean	31.1	12.3	3.8	0.35	0.13
Standard deviation	27.6	15.8	4.1	0.21	0.09
Maximum value	237.6	138.3	28.3	0.66	0.41

U_I = relative error on selection differential with no correction; R_I, P_I = relative error when correcting by Rawlings' and polynomial formulae, respectively; R_I^*, P_I^* = correction inefficiency rate by Rawlings' and polynomial formulae respectively.

Comparison with Meuwissen's formulae was possible on the 681 situations (out of the 1 383 simulated) with 1 or 2 intra-class correlation coefficients. These results are shown in table V. Average pairwise correlation coefficient was smaller than 0.5 for each situation with constant family size and 1 or 2 correlations (table Va). For these cases, Meuwissen's formulae were really better than ours: whereas Meuwissen's average error was smaller than 1% with a maximal error of 4%, the average error incurred with the polynomial formula was nearly 4 and 2% for 1 or 2 correlation cases, respectively. When family sizes were heterogeneous (table Vb), performances of our formula were maintained whereas those of Meuwissen's prediction, assuming a constant average family size, deteriorated and became worse than ours.

Fitting variances of selected observations

Only 606 situations of data set 1 (40 candidates) corresponding to $r \leqslant 0.6$ and constant family size were examined to adjust polynomial Q.

The results obtained suggested fitting $p - 0.5$ instead of p. Finally, polynomial Q was:

$$Q(p, \sigma_r) = \sigma_r^3(t_1\sigma_r + (p - 0.5)^2(t_2 + t_3\sigma_r)) \qquad [4]$$

with $t_1 = -13.8$ (0.7)
$t_2 = -265.8$ (9.8)
$t_3 = 558.0$ (24.7)

where the values in parentheses are the estimation standard errors. The R-square value of adjustment was found to be 0.84.

Table VI shows that large relative errors for variance of selected EBVs were observed on the simulated data. Considering candidates as independent led to an

Table V. Comparison of Meuwissen's, Rawlings' and polynomial formulae for selection differential according to family structure.

	U_I	M_I	R_I	P_I	M_I^*	R_I^*	P_I^*
a) *Constant family size*							
One-correlation case; 87 situations							
Mean	34.4	0.8	15.9	3.7	0.04	0.31	0.12
Standard deviation	46.8	0.8	28.4	5.5	0.04	0.24	0.09
Maximum value	237.6	4.0	138.3	28.1	0.24	0.85	0.46
Two-correlation case; 243 situations							
Mean	15.1	0.5	4.3	1.8	0.04	0.24	0.13
Standard deviation	11.2	0.5	5.8	2.0	0.04	0.18	0.10
Maximum value	64.5	2.3	31.1	12.7	0.24	0.81	0.41
b) *Variable family size*							
One-correlation case; 90 situations							
Mean	28.3	6.6	9.2	3.8	0.30	0.22	0.12
Standard deviation	22.5	7.7	14.6	4.9	0.20	0.20	0.09
Maximum value	101.0	39.5	71.5	19.8	0.40	0.73	0.35
Two-correlation case; 261 situations							
Mean	21.3	4.1	4.0	1.9	0.17	0.16	0.08
Standard deviation	12.9	6.1	4.8	2.5	0.21	0.14	0.07
Maximum value	71.5	8.5	28.9	14.6	1.23	0.70	0.40

U_I = relative error on selection differential with no correction; M_I, R_I, P_I = relative error when correcting by Meuwissen's, Rawling's and polynomial formulae, respectively; M_I^*, R_I^*, P_I^* = correction inefficiency rate by Meuwissen's, Rawlings' and polynomial formulae, respectively.

Table VI. Comparison of Owen and Steck's and polynomial formulae for variance of selected observations (606 situations: 40 candidates and constant family size).

	U_V	B_V	P_V	B_V^*	P_V^*
Mean	305	169	55	0.35	0.12
Standard deviation	879	521	205	0.22	0.14
Maximum value	8 850	4 282	1 686	0.84	1.20
% less than 10%	1.5	36.3	66.8	19.1	64.2
% less than 50%	36.0	68.6	88.4	75.9	98.0

U_V = relative error on variance with no correction; B_V, P_V = relative error when correcting variances by Owen and Steck's and polynomial formulae, respectively; B_V^*, P_V^* = correction inefficiency rate by Owen and Steck's and polynomial formulae, respectively.

average relative error equal to 305%. Correction attempts through Owen and Steck's or polynomial formulae decreased the amount of errors to 169 and 55%, respectively. On average, 88% of the error incurred with no correction for correlated EBVs was removed by our polynomial adjustment, whereas only 65% was removed by Owen and Steck's formula.

However, polynomial approximation for variances cannot be considered as safe as for selection differentials. Firstly, we were unable to find reasonable adjustment when data sets included variable family sizes. Secondly, in 2 cases out of the 606 analyzed, corresponding to 0.99 intra-class correlations, our correction led to errors higher than those incurred with no correction. Thirdly, the theoretical form of equation [2] does not preclude negative predictions.

Examination of values of Q according to p and σ_r showed that positive values of approximated variances are obtained for any selection rate as soon as σ_r is smaller than 0.35, and for selection rates higher than 2% for σ_r between 0.35 and 0.5. The polynomial approximation should not be used for σ_r greater than 0.5.

Cross-validation

The examples chosen correspond to situations where ignoring correlation between EBVs would lead to substantial relative errors: 10–20% for selection differentials and 30–130% for variances of the selected candidates (table VII). Rawlings' formula was found to be satisfactory (relative errors about 2%) when estimating selection differentials from moderate selection pressures (25–50%). Relative errors increased (6–12%) when selection was more severe (p around 2–4%). Owen and Steck's formula for predicting variances decreased biases but relative errors were still high (5–100%).

The efficiency of the polynomial formula was comparable to that of Rawlings' for moderate selection rates but was clearly superior for more severe selection, because relative errors by our formula in that case did not increase very much and were around 1–3%. Our polynomial formula for approximating variances was not entirely satisfactory but succeeded in giving better results than Owen and Steck's. The range of relative errors for variances was 0–50%; variances were overestimated for moderate selection pressures and underestimated for severe selection pressures.

DISCUSSION AND CONCLUSION

The objective of this work was to provide approximate expressions for selection differentials and corresponding variances on EBVs, easy to calculate and robust for any correlation structure between EBVs which are multinormally distributed. Hill (1977) proved that selection differential can be used to predict selection response when animals are ranked and selected on an optimum selection index.

Although no absolute proof can be given of the validity of our empirical approach for any situation, the moderate prediction errors observed on the calibration data sets (involving a very large diversity of situations) and on the cross-validation data sets (quite different from the former ones) lead one to think that these formulae are relatively robust and might be used for deterministic prediction on breeding schemes, especially when factorial mating designs are implemented and/or family sizes are variable (see, for instance, artificial insemination and natural service families).

However, particular situations should be addressed:

1) When r, the average pairwise correlation coefficient, is greater than 0.6, the situation corresponds to a population with all members closely related (for

Table VII. Relative errors on approximate selection differentials and variances for simplified breeding schemes (based on averages of 10 000 replicates per situation).

Breeding scheme	*Selection pressures*	*Selection differentials*				*Variances*			
		Observed	U_I	R_I	P_I	*Observed*	U_V	B_V	P_V
Beef cattle $h^2 = 0.25$ ($r = 0.176$; $\sigma_r = 0.246$)									
Female selection	46/93	0.725	10.4	0.2	1.8	0.282	27.7	5.2	0.1
Male selection	3/93	1.859	16.6	5.8	2.2	0.075	72.3	41.3	25.4
Beef cattle $h^2 = 0.10$ ($r = 0.223$; $\sigma_r = 0.309$)									
Female selection	46/93	0.698	14.6	1.0	2.4	0.253	42.5	10.1	3.2
Male selection	3/93	1.718	26.1	11.1	2.9	0.056	130.7	79.3	51.4
Dairy cattle $h^2 = 0.25$, adult MOET ($r = 0.160$; $\sigma_r = 0.220$)									
Daughter–dam selection	48/192	1.147	10.3	1.1	0.3	0.169	42.5	19.5	4.4
Sire–dam selection	4/192	2.015	16.9	7.0	0.6	0.057	108.8	75.4	8.7
Dairy cattle $h^2 = 0.25$, juvenile MOET ($r = 0.136$; $\sigma_r = 0.251$)									
Daughter–dam selection	12/48	1.144	10.6	1.4	0.6	0.170	36.1	17.4	6.0
Sire–dam selection	2/48	1.779	20.1	11.6	0.6	0.079	72.4	48.7	21.8

U_I, R_I, P_I = relative errors on selection differentials, with no correction, Rawlings' and polynomial approximations, respectively; U_V, R_V, P_V = relative errors on variances, with no correction, Owen and Steck's and polynomial approximations, respectively.

instance, a family with many full-sibs without own peformance information) and Rawlings' formula should be used.

2) With moderate selection pressures (p greater than 20%), although the polynomial approximation led to similar results, Rawlings' formula is recommended for the sake of simplicity.
3) With a hierarchical mating design leading to full-sibs nested within half-sibs, with families of constant size and constant intra-class correlation coefficients, Meuwissen's formula is preferred.

The situation is quite clear for variance corrections since our methods is much better than Owen and Steck's formula and the major part (88%) of the bias occurring with no correction is removed. However, the errors are still large. Important restrictions are that family size should be constant and that σ_r, the average standard deviation of the pairwise correlation coefficients involving a given candidate, should not exceed 0.5. For these reasons, further improvement should be investigated.

Additional heuristic research is needed to provide relevant approximations for selection differentials and variances of the selected candidates when variances are not constant (Perez-Enciso and Toro, 1991) and/or when EBVs of candidates do not have the same expectation, due to mixing of age cohorts, for instance. Such problems are very commonly encountered when attempting to implement deterministic predictions of genetic response.

REFERENCES

Burrows PM (1972) Expected selection differentials for directional selection. *Biometrics* 28, 1091-1100

Hill WG (1976) Order statistics of correlated variables and implications in genetic selection programs. *Biometrics* 32, 889-902

Hill WG (1977) Order statistics of correlated variables and implications in genetic selection programmes. II. Response to selection. *Biometrics* 33, 703-712

Meuwissen THE (1991) Reduction of selection differentials in finite populations with a nested full-half-sib family structure. *Biometrics* 47, 195-203

Nicholas FW, Smith C (1983) Increased rates of genetic change in dairy cattle by embryo transfer and splitting. *Anim Prod* 36, 341-353

Owen DB, Steck GP (1962) Moments of order statistics from the equicorrelated multivariate normal distribution. *Ann Math Stat* 33, 1286-1291

Perez-Enciso M, Toro MA (1991) A note on prediction of response to artificial selection with indices of unequal information. *Livest Prod Sci* 29, 335-340

Rawlings JO (1976) Order statistics for a special class of unequally correlated multinormal variates. *Biometrics* 32, 875-887

SAS/STAT User's Guide (1990) *The GLM procedure.* Vol 2, version 6, 4th edition, SAS Institute Inc, 892-996

Verrier E, Colleau JJ, Foulley JL (1991) Methods for predicting response to selection in small populations under additive genetic models: a review. *Livest Prod Sci* 29, 93-114

APPENDIX. ESTIMATING POLYNOMIAL REGRESSION COEFFICIENTS

Using the SAS procedure 'General Linear Models', dependent variate y (see text) corresponding to observed selection differentials for 1 383 situations, an R-square value equal to 0.86 was obtained from a polynomial of degree 5 involving the following predictive variates.

No	*Variate*	*Coefficient*
1	r	− 61.94
2	σ_r	60.15
3	$r\sigma_r$	− 332.25
4	r^2	319.75
5	σ_r^2	23.79
6	$r^2\sigma_r$	388.07
7	r^3	− 379.36
8	rp	481.17
9	$\sigma_r p$	− 440.61
10	$r\sigma_r p$	2 170.03
11	$r^2 p$	− 2 367.25
12	$r^2\sigma_r p$	− 2 528.51
13	$r^3 p$	2 756.11
14	$\sigma_r^3 p$	− 87.24
15	$r^3 p^2$	− 4 169.41
16	rp^2	− 748.60
17	$\sigma_r p^2$	664.25
18	$r\sigma_r p^2$	− 3 162.98
19	$r^2 p^2$	3 612.00
20	$\sigma_r r^2 p^2$	3 693.41

Each of these 20 coefficients was found to be significant (1‰ level). Coefficients were very similar but with opposite sign when examining coefficients of $r\,vs\,\sigma_r$, $r\sigma_r\,vs\,r^2$, $r^2\sigma_r\,vs\,r^3$, $rp\,vs\,\sigma_r p$, $r\sigma_r p\,vs\,r^2 p$, $r^2\sigma_r p\,vs\,r^3 p$, $rp^2 vs\,\sigma_r p^2$ and $r\sigma_r p^2\,vs\,r^2 p^2$. This suggested replacing the corresponding 16 original variates by 8 new variates $d(= r - \sigma_r)$, $rd, r^2 d, dp, rdp, r^2 dp$, dp^2, rdp^2, leading to a new polynomial with 12 coefficients $(8 + 4)$. Further factorization were performed twice on this polynomial. The final polynomial was that of our formula, with the corresponding final combined variates $d = r - \sigma_r$, $e = r(1 - r)$, $b = d(1 - 4.2e)$. The cost of such a simplification was moderate because the final R-square value was 0.85.