Evidence of a Bias in Genetic Evaluation due to Genomic Selection

Clotilde Patry[1,2] and Vincent Ducrocq[1]
[1]UMR 1313 INRA, Génétique Animale et Biologie Intégrative, 78 352 Jouy-en-Josas, France
[2]Union nationale des Coopératives d'Elevage et d'Insémination Animale (UNCEIA),
149, rue de Bercy, 75 595 Paris Cédex 12 France
email : clotilde.patry@jouy.inra.fr

Abstract

The inclusion of an extra selection step based on genomic evaluation in breeding schemes invalidates some of the assumptions leading to the optimal properties of BLUP. In particular, the selection process is no longer fully described in the analysis and the distribution of the Mendelian sampling term is no longer trivial. It is feared that the estimation of breeding values will be biased in national and international evaluations. The target of this study is to assess such a bias, measured as the difference between the average estimated and true breeding values, through simulations. True breeding values and estimated breeding value including genomic information are simulated. Considering different selection parameters, it was found that indeed the BLUP evaluation was biased. When genomic selection is implemented, BLUP solutions underestimate the true breeding values of young sires and their daughters. Further, reliabilities computed overestimate the squared correlation between true and estimated breeding values.

Keywords: (inter)national evaluation – selection bias – genomic selection - simulation

1. Introduction

In national and international genetic evaluations, BLUP applied to an animal model is commonly used to estimate animal breeding values. Under certain conditions, BLUP have properties yielding to an optimal efficiency of selection. Henderson (1982) focused on the assumptions required for populations that have undergone selection. Under normality, with the infinitesimal model, the mixed model equations of Henderson can account for selection if the process is based on a linear function of the observations and if all the data on which selection is based is included in the analysis.

However, if an extra selection step based on genomic information is implemented, these assumptions are no longer fulfilled. As selection is no longer based on observations, selection information will be missing in the analysis. An important consequence of culling animals without having corresponding observations in the evaluation is that the expectation of the Mendelian sampling effects is no longer 0 and its variance is no longer half the genetic variance As a result, assumptions leading to the nice properties of animal models are also no longer fulfilled (Kennedy *et al.*, 1988). Consequently, BLUP solutions may be biased at the national and international levels (e.g., Banos *et al.*, 2007; van der Beek, 2007). This would result in less accurate rankings of bulls and cows, and lower genetic progress.

The target of the present study is to assess based on simulations the selection bias introduced in estimated breeding values when a genomic preselection step is included.. The bias will be measured as the average difference between (simulated) true and estimated breeding values.

2. Materials and Method

2.1 Dataset:

The simulations were done following the structure of the national evaluation on a conformation trait, udder depth (UD), for the Holstein breed in France. Pedigree and performances are real so that initial BLUP solutions in the population are available. Selection bias will be assessed assuming a single trait selection on udder depth with a heritability of 0.36. Within the population, two groups of animals are particularly considered: (1) the candidate bulls which are represented by the cohort of young sires that have only one crop of daughters in the real data and (2) the cohort of daughters of these sires, for which records will be simulated. All the other animals will keep their "real" record as in the national evaluation. The simulations assuming genomic preselection consider that the candidate bulls sires have been pre-selected based on genomic information.

2.2 Principle:

The strategy is to compare BLUP solutions with true breeding values in two cases, one where young sires have undergone genomic selection (PS young sires) and the other not (REF young sires). First, estimated breeding values including genomic information (GEBV) and true breeding values (TBV) are simulated jointly for these young bulls. To mimic genomic selection, only GEBV-TBV sampled pairs for which GEBV is higher than a given truncation threshold are retained. Second, records from daughters of these young sires are simulated. Third, a BLUP evaluation using an animal model to estimate breeding values (EBV) on simulated data and real data (for the other categories of animals) is implemented. A sensitivity analysis is performed by varying some key parameters, leading to 8 scenarios to test. For each one, the computations described above are repeated 50 times.

2.3 Simulation of breeding values in the two contrasted populations of young sires

(a) Getting breeding values which include genomic information (GEBV)

In case of genomic selection, candidate bulls are evaluated before the preselection step without progeny records. The genomic evaluation combine the classical pedigree index with direct genomic values. This genomic information is considered to contribute as much as performances on n equivalent (additional) daughters. So, the (direct) genomic reliability of a sire S is R_{gen}:

$$\mathrm{R}_{gen} = \frac{n}{n+k} \text{ with } k = \frac{4}{h^2} - 1$$

Considering the information coming from pedigree (R_S), the reliability (R_{S+}) of young sire S after genomic evaluation is then:

$$R_{S+} = \frac{R_S + R_{gen} - 2R_S R_{gen}}{1 - R_S R_{gen}}$$

with $R_S = (R_{GS} + R_{GD})/4$, *GS* being the grand sire and *GD* the grand dam of S.

The genomic evaluation of sire S also has an impact on the reliability of its parents GS and GD. Let R_{GS+} be the reliability of the grand sire including the genomic information coming from its son (assuming GS has only one son, as an approximation):

$$R_{GS+} = \frac{R_{GS} + \frac{1}{4}R_{S+} - 2R_{GS}\frac{1}{4}R_{S+}}{1 - R_{GS}\frac{1}{4}R_{S+}}$$

Consider the EBV of a young sire as a random variable. Its distribution is:

$$\hat{a}_S \sim N(\frac{\hat{a}_{GS} + \hat{a}_{GD}}{2}, R\sigma_a{}^2)$$

Similarly, when the genomic information of *S* (R_{S+}) is combined with phenotypic information coming from the pedigree ($\hat{a}_{GS}$ and $\hat{a}_{GD}$), the genomic estimated breeding value (GEBV) of the young sire comes from:

$$\hat{a}_{S+} \sim N(\frac{\hat{a}_{GS} + \hat{a}_{GD}}{2}, R_{S+}\sigma_a{}^2)$$

However, to mimic genomic pre-selection, only the best sires are kept based on $\hat{a}_{S+}$. A threshold selection *t*, function of the selection intensity among candidates, has to be set. Only simulated values from animals, whose GEBV $\hat{a}_{S+}$ is larger than t, will be kept.

For the *GS* and *GD*, classical information from phenotypic data is combined with the genomic information from S to get their GEBV $\hat{a}_{GS+}$ and $\hat{a}_{GD+}$.

(b) Simulation of the true breeding values in the PS and REF populations

All the breeding values related to a sire (EBV, GEBV and TBV of a young sire and its parents) are jointly normally distributed. In fact, for each S, $\hat{a}_{GS}$ and $\hat{a}_{GD}$ are assumed known as their value come from the real data set analysis. The distribution of the other TBV and GEBV can be obtained conditionally on these known parents' EBV. We have:

$$\begin{bmatrix} a_{GS} \\ a_{GD} \\ a_S \\ \hat{a}_{GS+} \\ \hat{a}_{GD+} \\ \hat{a}_{S+} \end{bmatrix} \sim N \left[\begin{pmatrix} \hat{a}_{GS} \\ \hat{a}_{GD} \\ \frac{\hat{a}_{GS}+\hat{a}_{GD}}{2} \\ \hat{a}_{GS} \\ \hat{a}_{GD} \\ \frac{\hat{a}_{GS}+\hat{a}_{GD}}{2} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \lambda_{GS} & 0 & \frac{1}{2}\lambda_{GS} & \Delta_{GS} & 0 & \frac{1}{2}\Delta_{GS} \\ 0 & \lambda_{GD} & \frac{1}{2}\lambda_{GD} & 0 & \Delta_{GD} & \frac{1}{2}\Delta_{GD} \\ \frac{1}{2}\lambda_{GS} & \frac{1}{2}\lambda_{GD} & \lambda_{PED} & \frac{1}{2}\Delta_{GS} & \frac{1}{2}\Delta_{GD} & R_{S+}-R_{PED} \\ \Delta_{GS} & 0 & \frac{1}{2}\Delta_{GS} & \Delta_{GS} & 0 & \frac{1}{2}\Delta_{GS} \\ 0 & \Delta_{GD} & \frac{1}{2}\Delta_{GD} & 0 & \Delta_{GD} & \frac{1}{2}\Delta_{GD} \\ \frac{1}{2}\Delta_{GS} & \frac{1}{2}\Delta_{GD} & R_{S+}-R_{PED} & \frac{1}{2}\Delta_{GS} & \frac{1}{2}\Delta_{GD} & R_{S+}-R_{PED} \end{pmatrix} \sigma_a^2 \right]$$

with $\lambda = 1 - R$, $\Delta = R_+ - R$ and $R_{PED} = \dfrac{R_{GS} + R_{GD}}{4}$.

For any parent, the first set of values (a_{GS}, $\hat{a}_{GS}$) drawn from the defined distribution is kept later on for any other progeny.

In the reference population, there is no condition on the GEBV of sires (= no selection, t = -inf). In contrast, in the PS population, the GEBV of a young sire has to fulfil the threshold condition. Many samplings are required, their number depending on the selection intensity. Once the condition is fulfilled, the true breeding value, a_S computed together with $\hat{a}_S$, is kept.

2.4 Simulation of daughters' records:

Consider the single trait animal model:

$$y_i{*} = HRC + a_i + e_i$$

where y_i^* are the UD records already corrected for fixed effects (age at calving and stage of lactation). Again, the same data structure as in the initial population is assumed. The estimate of HRC (herd-round-classifier) computed in the initial evaluation is used to simulate y_i^*.

The infinitesimal genetic model implies that any sire's daughter breeding value is the sum of half of the sum of the parental breeding values and a Mendelian sampling term (φ). Already knowing the sire breeding value (a_S), the other variables are drawn from normal distributions:

$$a_D \sim N(\hat{a},(1-R_D\sigma_a^2))$$

for the dam breeding value, $\varphi_i \sim N(0,\frac{1}{2}\sigma_a^2)$ and e$_i$~ $N(0,\sigma_e^2)$

Hence, for each secanrio (with or without genomic preseletion of the young sires), a different set of daughters' records is obtained. The performance records of the other animals in the whole population are kept unchanged.

2.5 Comparing evaluations:

Finally, two BLUP evaluations are run separately in the REF and PS populations using an in house software "Genekit" (Ducrocq V., 2006). EBV are obtained for all candidates. Average difference between TBV and EBV are then computed for each cohort of young sires and their daughters.

2.6 Sensitivity analysis:

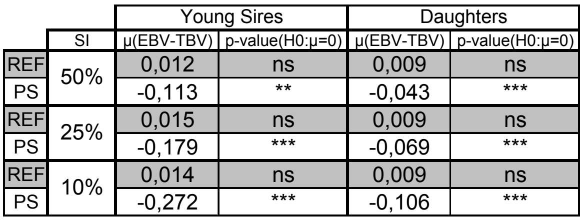
To assess the importance of selection bias and to better understand its variation, different scenarios were considered. They are based on the combinations of different selection criteria (selection based on GEBV or on estimated Mendelian sampling term), different levels of selection intensity (SI=50, 25 and 10%) and levels of accuracy of the genomic evaluation (genomic equivalent daughter contributions n assumed equal to 10 or 20). This leads to 8 scenarios. For each one, 50 replicates were simulated.

3. Results

(a) Underestimation of breeding values:

Whatever the selection criterion, the selection intensity or the group of interest (young sires or their daughters), BLUP solutions in the PS populations were found to be biased, in contrast with what was observed in the REF populations: when a genomic selection was implemented, the difference between estimated breeding values and true breeding values were significantly different from zero. Further, the contrast is negative: BLUP estimated breeding values of pre-selected animals and their progeny were underestimates of the true breeding values.

Table 1. Bias in REF and PS populations when pre-selection is based on GEBV (in standard genetic deviation).



		Young Sires		Daughters	
	SI	μ(EBV-TBV)	p-value(H0:μ=0)	μ(EBV-TBV)	p-value(H0:μ=0)
REF	50%	0,012	ns	0,009	ns
PS		-0,113	**	-0,043	***
REF	25%	0,015	ns	0,009	ns
PS		-0,179	***	-0,069	***
REF	10%	0,014	ns	0,009	ns
PS		-0,272	***	-0,106	***

The magnitude of the bias increased with the selection intensity. In the cohort of young sires, the mean bias values over the 50 repetitions ranged from -0.11 (with SI=50%) to -0.27 (SI=10%) genetic standard deviation of the udder depth trait. The mean standard deviation of the bias was equal to 0.31 for all the selection intensity values. In the young sires cohort, the largest bias for a sire could be as large as one genetic standard deviation for udder depth (from -0.93 (SI=50%) to 1.09 (SI=10%)).

In the cohort of daughters, the difference between true and estimated breeding values was smaller. Values ranges from -0.04 (SI=50%) to -0.11 (SI=25%). However, the standard deviation of the bias was much larger than in the young sires cohort (0.73 whatever the selection intensity).

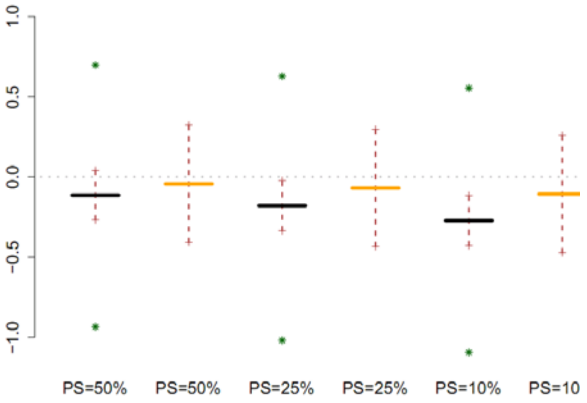


Figure 1. Biais magnitude in the cohorts of young sires (dark lines) and their (light lines) daughters for increasing selection intensity (PS) – *dotted lines = bias standard deviation, stars=maximum and minimum values of bias in a set of sires.*

Consider now a genomic preselection of young bulls based on the estimated mendelian sampling term including genomic information. In such a case, the animal model assumption of null expectation of mendelian terms is even more clearly incorrect. Bias was found to be larger: when selection intensity was 25%, the mean bias was -0.31 *vs* -0.18 in the case of selection based on EBV (figure 3).

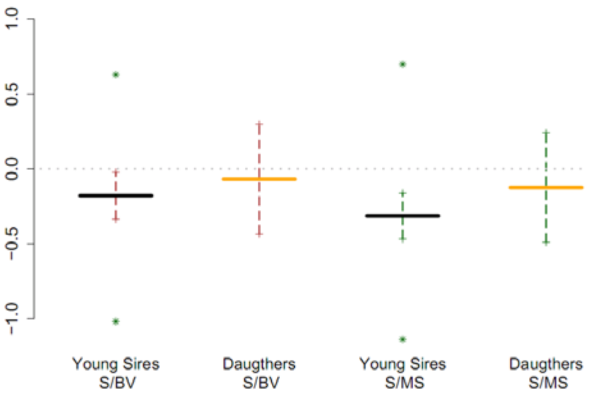


Figure 2. Bias when selection is based on EBV (S/BV) vs on Mendelian sampling term (S/MS) – ***dotted lines = bias standard deviation, stars=maximum and minimum values of bias in a set of sires.***

(b) <u>Reliabilities</u>

By definition, the reliability of an evaluation is the square correlation between true and estimated breeding values. Because we use simulation, this theoretical reliability can be computed and compared to the one obtained from the mixed model equations. Whatever the selection criterion, BLUP reliabilities were overestimates of the real ones (table 2). The average true reliabilities ranged from 0.71 (SI=10%) to 0.81 (SI=50%), whereas the one based on the data structure was equal to 0.90. Note that reliabilities were less altered when animals were selected based on their Mendelian sampling term. In the same way, reliabilities were less overestimated in the daughters' cohort.

Table 2. Reliabilities computed as the squared correlations between TBV and EBV vs reliabilities based on BLUP mixed model equations.

			Young Sires		Daughters	
		SI	CORR²(TBV,EBV)	R²(BLUP)	CORR²(TBV,EBV)	R²(BLUP)
S/BV	REF	50%	0,88	0,90	0,44	0,51
	PS		0,79	0,90	0,37	0,51
	REF	25%	0,88	0,90	0,43	0,51
	PS		0,73	0,90	0,35	0,51
	REF	10%	0,88	0,90	0,43	0,51
	PS		0,70	0,90	0,34	0,51
S/MS	REF	25%	0,88	0,90	0,43	0,51
	PS		0,83	0,90	0,39	0,51

Discussion and Conclusion

This study considers different genomic selection intensities of young sires and measures for different cohorts the impact of pre-selection on GEBV. It was showed that, as feared, a bias due to genomic pre-selection exists: the EBV of pre-selected bulls and their daughters are underestimated, sometimes to a large extent (up to one genetic standard deviation) and standard reliabilities are overestimated. As a result, bull rankings are less accurate and genetic progress will be lower. Only a simple situation, likely to occur in a near future, was considered here. It is likely that this bias may increase with time and that its sign may become unpredictable, when a variable proportion of herdmates of daughters of young bulls are also daughters of "genomic pre-selected bulls". The classical evaluation tools will no longer be adapted and ways to correct this bias must be found. See Patry and Ducrocq (2009) for potential directions for improvement.

This study considered only the case of national evaluations. The challenge of correcting pre-selection bias is even more relevant when international evaluations are considered, especially when selection practices and availability of pre-selection information reliabilities strongly vary between countries.

References

Banos, G. *et al.* 2007. Interbull Scientific Advisory Committee annual report 2007, Dublin, Ireland. *www-interbull.slu.se*

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics 31,* 423-447.

Kennedy, B.W. et al. 1998. Genetic properties of animal models. *J.Dairy Sci. 71(suppl2),*17-26.

Patry, C. & Ducrocq, V. 2009. Bias due to genomic selection. *Interbull Bulletin 39,* 77-82.

Van der Beek, S. 2007. Effect of genomic selection on national and international genetic evaluations. *Interbull Bulletin 37,* 115-118.