Proceedings, 10th World Congress of Genetics Applied to Livestock Production

New simulation method to create data sets with a desired genetic trend

A.- M. Tyrisevä[1], M. H. Lidauer[1], G. P. Aamand[2] and E. A. Mäntysaari[1]
[1]MTT Agrifood Research Finland, Jokioinen, Finland, [2]NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark.

ABSTRACT: The simulation method consists of two main steps. First, original vector of observations is replaced with a vector of yearly increasing values that are used to predict pseudo breeding values (BVs) carrying a desired genetic trend. The pseudo BVs and the pedigree information are then used to calculate Mendelian sampling (MS) terms for each animal. Finally, the simulated data sets carrying the genetic trend are generated utilizing the MS terms created in the first step. The method retains the original structure of the pedigree and any number of random and fixed effects can be fitted. The method is shown to yield data sets expressing closely the targeted genetic trend. The method was used to study the effect of genomic pre-selection on BVs and MS terms. The results revealed a clear bias in the estimated BVs and MS terms in bulls after genomic pre-selection.
Keywords: evaluation bias; simulation of genetic trend; genomic pre-selection

Introduction

Simulation studies are inseparable part of the methodological development in animal breeding. By using simulated data sets, true breeding values (BVs) are obtained and the accuracy of the methods under development can be evaluated by comparing the true BVs with the estimated ones. The easiest way to simulate test data sets is to generate a simple family structure with a constant number of sires and progenies per sire. However, real populations are far more complex and in many cases this simplistic scheme is not the best option. Instead, data structures that better corresponds the reality are needed.

Another way to simulate data sets is to utilize a structure of the real population and a suitable evaluation model and to simulate observations according to the random effects of the model by using, e.g., Monte Carlo sampling procedures (Meyer (2002); Lidauer et. al (2011)). However, if it is desired that the evaluation model detects a genetic trend, the structure of the data should be such that selection can be distinguished from it. For this kind of research problems the Monte Carlo methods exploiting real data structures cannot be used.

Selection can be simulated in the data by simulating a full breeding program (e.g., Lillehammer et al. (2011)). In that case, generations are simulated one by one and selection is practiced before generating a new generation. However, this procedure does not retain the original data structure. The aim of this study was to develop an approximative simulation method that creates a genetic trend in the data, but retains the original structure of the pedigree and allows inclusion of any number of random and fixed effects. By using this method, we studied how the genomic pre-selection not accounted for by the evaluation model affects breeding values and Mendelian sampling terms.

Materials and Methods

Method. The dependent variable in the original data is replaced with the observations having a desired annual trend. From this data set, pseudo BVs expressing a genetic trend, are predicted and they become synchronized with parent and progeny averages, and the expected yearly means of BVs. BVs are then used together with the pedigree information to obtain Mendelian sampling (MS) terms for each animal. Hence, the genetic trend is transmitted in the MS terms. To ensure that the MS terms of parents would not be regressed towards yearly means, observations of parental animals in this first step are set missing. In the next step, true BVs are generated recursively generation by generation so that the BV of each animal consists of the parental mean, of the random MS term sampled from the normal distribution, and of the MS term created in the first step and carrying the genetic trend. Finally, the other random effects in the model, including residual, are generated and the generated random effects are summed to form observations.

Normally, the MS variance can be expressed as $\sigma_\varphi^2 = d_{jj}\sigma_u^2$, in which d_{jj} is the diagonal of an animal j in the decomposition of $\mathbf{A}=\mathbf{LDL}^\mathbf{T}$, $\mathbf{L}$ is the lower unitriangular transition matrix**,** and σ_u^2 is the additive genetic variance. Under the non-zero expectation of the MS terms, this does not hold true anymore. The MS variance increases with the variance of the MS terms created in the first step, leading also into inflated variance of BVs. To avoid this, a variance correction can be carried out:$\sigma_\varphi^2 = (1-k)d_{jj}\sigma_u^2$. According to the standard formula by Falconer and Mackay (1996), $k = i(i-x)$ where i is the selection intensity and x is the deviation of truncation point from the mean in standard deviation units. The selection intensity can be further formulated as $i = E[\varphi]/\sigma_\varphi$, in which $E[\varphi]$ is the expected value of the MS term. Since $(1-k)$is an exponential function of i, a satisfactory approximation can be obtained by a linear fit on its logarithmic value. A good fit was obtained with the formula:

$$(1-k)_j = Exp(-1.18969|i| + 0.10805i^2),$$

where $|i|$ is the absolute value of *i*.

Example data. To test the method, a field data set of 754 600 Danish Holstein cows from 2000 herds was sampled. The time interval covered 20 years and the pedigree information included 1.2 million animals. Only the herd and the pedigree structure were retained from the original data and an artificial trait was simulated. One record was generated for each cow. The model used in both steps included a fixed herd effect and random additive genetic and residual effects. In calculation of the pseudo BVs, a heritability of 0.05 was used to ensure that the base level of the MS means in each birth year class remained zero. When the final data sets were simulated, the heritability of 0.25 was used. To mimic protein production a genetic standard deviation (SD) was assumed 41.

Design of the study. First, a genetic trend of 15% of the genetic SD was created by the data used to solve pseudo BVs and to calculate MS terms. This was done only once and used for all data replicates thereafter. Second, control and genomic pre-selection (GPS) schemes with 50 data replicates for each were created. The same seeds were used for replicates from the control and GPS schemes. To create a genomic pre-selection in the GPS scheme, all bulls from the birth year class 2000 onwards were assumed to be genomically pre-selected. The MS terms obtained from the first step for these bulls were increased with the $MS+$ term that was calculated as:

$$MS+ = \sigma_\varphi \times i \times r^2 = \sqrt{1650/2} \times 1.755 \times 0.60 = 31,$$

in which r^2describes the uncertainty of the predictions. The $MS+$ corresponds to the selection of the best 10% of the genomically tested bull calves. The MS terms of the cows were unaltered.

Analyses. True and estimated BVs were obtained under both schemes and used to calculate within-year means of BVs. The simulations were carried out with MiX99 that was modified to generate BVs from the MS term distribution with the non-zero expectation (Lidauer et al. 2011). A bias expressed as an estimated BV – true BV was also calculated, as well as a ranking of the top 10% of the bulls encompassing the birth years around the start of GPS. The MS terms from the true and estimated BVs were obtained from the program developed for the validation of the MS trend (Tyrisevä et al. 2012) and from them within-year means were calculated.

Results and Discussion

Within-year means of BVs and the estimation bias for bulls are shown in Table 1. The means of pseudo BVs came from one data set, all other results were averaged over 50 data replicates. The genetic trends of pseudo and true BVs from the control scheme were very similar, with the same yearly genetic progress (7.1 units / year). Further, the within-year means of BVs estimated from the generated data sets followed the underlying true BVs very closely, with the yearly genetic progress of 7.0 units / year. A steep increase in the level of true BVs was observed, when the GPS started in 2000 (Table 1, Figure 1). The genetic progress was clearly higher for the true BVs under the GPS scheme than for the control scheme, being 9.9 units / year. Only part of this genetic progress could be detected, when the BVs were estimated from the generated data sets: the genetic progress was 8.5 units / year. The observed bias under the GPS scheme was twice as high as that in the control scheme from the beginning of the studied time interval, increased steadily until the start of the GPS and after that increased steeply due to under-estimation of the BVs. A slight increase in the yearly means of true and estimated BVs could also be detected in cows two years after the start of GPS in bulls, when the first GPS bulls became sires (Figure 1). The results are in a good accordance with those obtained by Patry and Ducrocq (2011).

Table 1: Within-year means of pseudo, true and estimated (Est) breeding values (bv) for bulls, as well as average biases expressed as estimated bv - true bv. Results from control and genomic pre-selection (GPS) schemes were averaged over 50 replicates. GPS started in 2000.

Year	Pseudo	Control			GPS		
		True	Est	Bias	True	Est	Bias
1990	-18.76	-18.63	-17.89	0.74	-18.63	-17.32	1.31
1991	-6.14	-5.38	-4.84	0.54	-5.38	-4.17	1.22
1992	0.15	0.50	1.31	0.81	0.50	2.02	1.52
1993	7.23	7.25	8.14	0.90	7.25	9.02	1.78
1994	14.54	14.89	15.12	0.23	14.89	16.61	1.71
1995	21.53	21.62	22.24	0.62	21.62	23.58	1.95
1996	31.28	31.44	31.82	0.38	31.44	33.55	2.11
1997	37.85	38.45	38.60	0.15	38.45	40.73	2.28
1998	45.87	45.98	46.15	0.18	45.98	48.55	2.58
1999	48.51	49.18	49.21	0.04	49.18	51.89	2.71
2000	56.39	56.63	56.55	-0.08	87.73	72.31	-15.42
2001	64.67	65.39	65.22	-0.17	96.24	81.60	-14.64
2002	70.05	70.56	70.10	-0.46	101.79	87.93	-13.87
2003	75.16	75.64	74.89	-0.76	107.06	93.66	-13.41

Both true and estimated MS means were very close to zero in bulls under the control scheme (Figure 2). The true MS means under the GPS scheme were identical with the MS means from the control scheme until the start of GPS. After that, an expected rise of +31 could be observed. The estimated MS means clearly deviated from zero after start of GPS, but the increase in mean was only around 1/3 of that seen in the true MS means. The true and estimated MS means under the control and GPS schemes were in practice zero in cows (Figure 2). Patry and Ducrocq (2011) found similar results from their simulations.

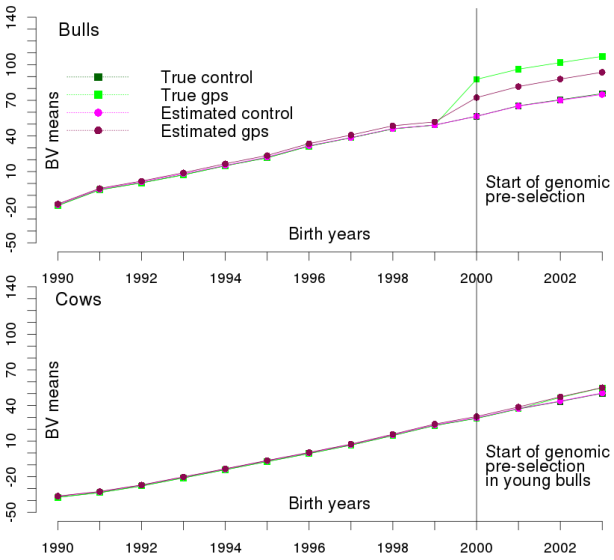

Figure 1: Within-year means of true and estimated breeding values in bulls and cows from control and genomic pre-selection (gps) schemes. Means were averaged over 50 replicates.

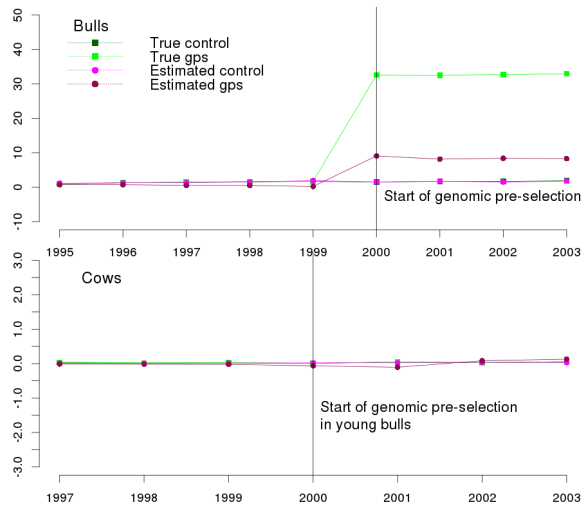

Figure 2: Within-year means of true and estimated Mendelian sampling terms in bulls and cows from control and genomic pre-selection (gps) schemes. Means were averaged over 50 replicates.

Proportion of the birth year classes among top 10% of the bulls based on true and estimated BVs under the control and GPS schemes are collected in Table 2. Only the years encompassing the start of GPS – from 1997 to 2002 – were studied. Differences in ranking based on true and estimated BVs under the control scheme were minor, whereas larger differences were observed under the GPS scheme (Table 2). In the GPS scheme, compared to the ranking based on true BVs, the birth year classes from 2000 to 2002 were under-represented when ranking was based on the estimated BVs. When bulls were ranked according to the true BVs, the birth year classes before year 2000 provided 34.5% and 12.3% of the top ranking bulls in the control and GPS schemes, respectively. With estimated BVs, the proportion remained almost unchanged (34%) in the control scheme, but in the GPS scheme the proportion became clearly too large (14.9%). It is possible that the old bulls having second crop daughters became over-estimated.

Table 2: Distribution of the top 10% of the bulls into birth year classes 1997-2002 according to true and estimated (Est) breeding values from control and genomic pre-selection (GPS) schemes. Results were averaged over 50 replicates. GPS started in 2000.

Years	Control		GPS	
	True	Est	True	Est
2002	24.5	24.2	33.6	31.1
2001	26.8	27.4	36.3	35.3
2000	14.3	14.4	17.8	18.6
1999	12.5	12.5	4.6	5.7
1998	14.2	14.7	5.3	6.6
1997	7.8	6.8	2.4	2.6

Conclusion

A simple method was developed to generate simulated data sets with genetic trend. In this study, the existing genetic evaluation software was modified to generate BVs from the MS term distribution with the non-zero expectation. The corresponding modifications are easy to implement to any genetic evaluation software. The method retains the original structure of the pedigree and any number of random and fixed effects can be generated. The method was used to illustrate the effect of genomic pre-selection on the BVs and MS terms. The targeted genetic trend was precisely transmitted in the simulated data sets. The results revealed a clear under-estimation of the breeding values in bulls after the start of genomic pre-selection, as well as a notable deviation from zero both in true and estimated MS means. Results for cows were only slightly affected. Bulls born after start of genomic pre-selection were under-represented in the top ranking, when the ranking was based on estimated BVs.

Literature Cited

Falconer, D. S. and Mackay, T. F. C. (1996). Introduction to quantitative genetics, fourth ed. Longman, Essex.

Lidauer, M. H., Matilainen, K., Mäntysaari, E. A. and Strandén, I. (2011). MiX99, release IV/2011.

Lillehammer, M., Meuwissen, T. H. E. and Sonesson, A. K. (2011). J. Dairy Sci., 94: 493-500.

Meyer, K. (2002). Proc 7th WCGALP, Communication No. 28-27.

Patry, C. and Ducrocq, V. (2011). J. Dairy Sci., 94:1011-1020.

Tyrisevä, A. - M., Mäntysaari, E. A., Jakobsen, J. et al. (2012). Interbull Bull., 46: 97-102.