Estimation of GEBVs Using Deregressed Individual Cow Breeding Values

Esa A. Mäntysaari*[1]*, Minna Koivula*[1]*, Ismo Strandén*[1]*,
***Jukka Pösö*[2] *and Gert P. Aamand*[3]**
[1]*Genetic Research, MTT Agrifood Research Finland, Jokioinen,*
[2]*Faba Coop, Vantaa, Finland,*
[3]*NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark.*

Abstract

Individual cow breeding values were deregressed using animal model and full national pedigree with 4.579 million animals. The deregressed proofs (DRPs) were used to recalculate the EBVs for their sires. The correlation between recalculated EBVs and original EBVs were 0.9997-1.0000, with small differences in EBVs of bulls without own daughters. Finally, the animal model DRPs were used to solve single-step genomic evaluations utilizing genotypes of 4725 bulls. The single-step approach was implemented using PCG algorithm and iteration on data. The Interbull GEBV test validation reliabilities for milk, protein and fat GEBVs were 0.35, 0.36 and 0.45, respectively. These were on average 0.03 higher than the single-step GEBVs with sire model DRPs. Overall the animal model deregression was computationally inexpensive. Also the single-step approach with 3.40 million cow records and three traits simultaneously took less than 1 hour wall clock time.

Introduction

Most genomic evaluations are currently based on two stage- approach. First, the genomic model is fitted to genotyped reference animals that have known “phenotypic records”, i.e., daughter based EBVs. Second, the genomic model is used to predict direct genomic values (DGV) of candidate animals without own records. Frequently DGV and the EBV based on pedigree index are combined together before publishing (GEBV; genomic enhanced breeding value). Combining can be based on selection index (VanRaden *et al.*, 2008), pseudo records (Ducroucq and Liu 2009, or bivariate BLUP (Mäntysaari and Strandén, 2011). Alternatively, single-step method can be used. In the single-step analysis the phenotypic records are combined directly with genomic information, and the resulting GEBV already combine both sources of information optimally (Christensen and Lund, 2010; Misztal *et al.*, 2009; Aguilar *et al.*, 2010).

Typical phenotypic records in genomic evaluations are either daughter yield deviations (DYDs), or deregressed bull EBVs (DRP). The DRPs are more popular because of their easy availability. Especially when reference population is based on genotypes from different countries, the DYDs are difficult to obtain. The single-step approach enables use of original phenotypic records, but computational reasons can favor the use of DYDs or DRPs. Bull DRPs are computed using the bull EBVs and the pedigrees of the bulls. Bull dam records (i.e. EBVs) are not needed. However, in the DGV-EBV combining step, the pedigree indices can include dam EBVs (vanRaden *et al.,* 2008) or it can be excluded because of anticipated bull dam evaluation bias (Reinhard *et al.,* 2009).

Alternatively to bull DRPs the deregression can be based on EBVs for the cows only. Resulting animal model DRPs are appealing choice for bivariate blending but also for single-step genomic evaluations. They contain all genetic information available in data, but are much easier to handle than the original data records. For example the full input data file of Nordic NAV TD model evaluation has size of 2990 Mb, while the file with three EBVs (milk, protein and fat) and their corresponding reliabilities for 3.40 million cows needs 208 Mb.

The objective of this study was to test the feasibility of deregression of individual cow EBVs. The animal model DRPs were first tested for reversibility by recalculating the bull EBVs from daughter DRPs. And finally the animal model DRPs were used to compute single-step genomic evaluations. The analyses were done with the full Nordic Red Cattle (RDC) population.

Material and Methods

The NAV evaluations for yield assume 27 different test day traits (Lidauer *et al.*, 2006). For this study, composite milk, protein and fat EBVs and their corresponding effective daughter contributions (EDC) were used. For the cows EDC_{cow} was defined as in Interbull (2004), with exception that dam reliability was not added to daughter reliability. The EDC_{cow} were calculated by the APAX99 -program (Strandén *et al.*, 2001) for the 3.40 million cows with records. The variance parameters in the EDC approximation were for the average daily TD (Interbull 2004), and the same values (h^2_{milk}=0.40, $h^2_{protein}$=0.28, and h^2_{fat}=0.32, and all r_g and r_e equal to 0.0) were used throughout this study. The deregression was done using Secant method in option *deregress* in MiX99 (Strandén and Mäntysaari, 2010). Deregression used the full pedigree of 4.58 million animals in NAV evaluation. The sires were only in the pedigree, and their EDCs or EBVs were not used. The reversibility of deregression was tested by recalculating EBVs from the DRPs.

The DRPs were used to compute GEBVs for all animals in the pedigree with the single-step approach (Christensen and Lund, 2010; Aguilar *et al.*, 2010). The single-step evaluations were implemented with PCG iteration on data in MiX99. Relationships based on numerator relationship matrix ($\mathbf{A}^{-1}$) were formed by reading the pedigree file, and the block of $\mathbf{H}^{22}$ ($\mathbf{G}^{-1}$- $\mathbf{A}_{22}^{-1}$) pertaining to genotyped animals (Aguilar *et al.*, 2010) was read (from IO-cache) within each iteration round. Genotypes for 37996 SNPs were used for 4725 bulls with progeny.

The **G** matrix was a scaled version of method 1 matrix in VanRaden *et al.* (2008). The matrix was scaled by dividing it by a scalar in order to have on average same diagonals as A_{22}, and regressed 10% towards $A_{22.}$ The regression can be interpreted as a fraction of genetic variance not explained by SNP genotypes. The value 10% had been found optimal in earlier sire model single- step analysis (Koivula *et al.*, 2011). The $\mathbf{A}_{22}$ was constructed with RelaX2 (Strandén and Vuori, 2006) using the full pedigree and algorithm suggested by Collaeu (2002). To assure the correctness of the genomic model, variance components were estimated from the DRPs using sire model and **G** matrix for genotyped bulls. Two alternative single-step evaluations were tested: single-step with deregression variance parameters, and single-step with estimated variance components.

GEBVs were validated with Interbull GEBV test (Mäntysaari *et al.*, 2010). First the GEBVs for the test bulls were calculated from a reduced data where the daughters of the test bulls had been removed. The 809 test bulls were chosen using data truncation (Koivula *et al.*, 2011), so that they had no daughters in 2005 NAV evaluations but had >= 20 daughters in 2010 evaluations. The number of daughters removed was 153,386. Second, the GEBVs of the test bulls were used to predict the DRPs of the bulls calculated with usual sire model deregression.

Results

The deregression for all three traits simultaneously took 13 min 38 seconds in 2.8 GHz Xenon™ CPU. Figure 1a shows the means of the milk EBVs for cows by birth year in Finland, Sweden and Denmark, and Figure 1b the same for DRPs. As expected the trends were alike. Figure 1c shows the yearly standard deviations of milk DRPs and EBVs. The SD of the DRPs was roughly 3 times larger than that for EBVs.

Refitting a three trait animal model BLUP to the cow DRPs took 9 min 56 seconds. The correlation between the original EBV and the recalculated EBVs for bulls from different countries were 0.9997 to 0.9999 depending on traits and bull's country of registration. In Figure 2a the recalculated protein EBV for Finnish bulls is plotted against original EBVs. There were some discrepancies in EBVs for the older bulls that have no daughters. Figure 2b shows only bulls that had more than 20 daughters. With this limitation the correlations between original and recalculated EBVs were 0.9999-1.0000.

The variance components estimated from the cow DRPs differed only slightly from the parameters used in deregression. The “genomic heritabilities” were 0.42, 0.41 and 0.30 for milk, protein and fat, respectively. The computational needs for the single-step analysis varied depending on variance components (parameters or data estimated), but generally the three trait model converged with 850-1180 iterations and in about 40-50 minutes. Tsuruta *et al.* (2011) implemented single-step approach on US Holstein type trait data with a similar size to ours. Based on validation reliability they suggested that solutions do not change meaningfully after the convergence criteria reaches 10^{-14}. In MiX99 this would correspond criteria Cr < 10^{-7}, which was reached at round 445, and the computing time for 3 traits would be under 20 minutes.

The model validation results are in Table 1. Both genomic models gave close to same R^2 validation reliabilities 0.35, 0.36-0.38 and 0.45 for milk, protein and fat, respectively. The parent average based on same data and animal model but without genomic information gave 0.12 lower reliabilities for milk and protein, but 0.17 lower for fat. In all the traits the b_1 regression coefficient values were significantly lower than the expected value of one indicating that differences among bulls were over predicted by GEBV.

For comparison purposes Table 2 lists the GEBV test results from 2-step approach and from sire model single-step genomic evaluations (Koivula *et al.*, 2011). The results are computed using the same candidate sires, and are therefore closely comparable to results in the Table 1. On average the animal model GEBVs had 0.03 higher validation R^2 than sire DRP single-step GEBVs, and GEBVs based on sire DRPs were again 0.03 better than DGVs based on sire DRPs. Unfortunately, information in GEBVs tend to be more inflated than in DGVs. This doesn't seem to be related with pedigree information, because it is clearer with sire-mgs PA (Table 2) than with sire-dam PA. We also remind that the dam DRPs of the bull dams, as used here, are not as biased as the EBVs of the young bulls in practice, because the EBVs of their sons have reduced the overestimation. In 2-step evaluations the variation in DGVs can be down-scaled to reflect their average accuracy. This reduces the upward bias. However, scaling is not possible in single- step GEBVs unless they are separately published only for young bulls without own daughters.

Conclusions

Animal model deregression of individual cow records seems to work well. It is computationally feasible even for large populations. Deregression can be easier than computing yield deviations, especially when the evaluations are based on complicated multi-trait or random regression models. Deregressed daughter records seem to work well in single-step genomic evaluations. It automatically includes some of the bull dam information into GEBVs although this might increase the bias in young bulls which have non-genotyped dams. The single-step iteration can be effectively implemented using iteration on data.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci. 93*, 743-752.

Christensen, O.F. & Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol. 42*, 2.

Colleau, J.J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol. 34,* 409–421.

Ducrocq, V. & Liu, Z. 2009. Combining genomic and classical information in national BLUP evaluations. Proc. Interbull meeting 21.-24.8.2009. Barcelona, Spain. *Interbull Bulletin 40*, 172-177.

Interbull. 2004. Code of Practice April 27th 2004. http://www.interbull.se/

Interbull. 2010. Interbull validation test for genomic evaluations – GEBV test. June 2010. http://www.interbull.org/images/stories/GEBV_validationtest_June2010.pdf. Accessed 10.5.2011.

Koivula, M., Strandén, I. & Mäntysaari, E.A. 2011. Comparison of different methods to calculate genomic predictions – results from SNP-BLUP, G-BLUP and one-step H-BLUP. EAAP 29.8-2.9.2011, Stavanger, Norway.

Lidauer, M., Pedersen, J., Pösö, J., Mäntysaari, E.A., Strandén, I., Madsen, P., Nielsen, U.S., Eriksson, J.-Å., Johansson, K. & Aamand, G.P. 2006. Joint Nordic Testday Model: Evaluation Model. Proc Interbull June 4-6, 2006, Kuopio, Finland. *Interbull Bulletin 35,* 103-107.

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci. 92*, 4648-4655.

Mäntysaari, E.A., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. Proceedings of the Interbull International Workshop March 4-5, 2010, Paris, France. *Interbull Bulletin 41*, 17-21.

Reinhardt, R., Liu, Z., Seefried, F. & Thaller, G. 2009. Implementation of genomic evaluation in German Holsteins. Proc. Interbull meeting 21.-24.8.2009. Barcelona, Spain. *Interbull Bulletin 40,* 219-226.

Strandén, I., Lidauer, M., Mäntysaari, E.A. & Pösö, J. 2001. Calculation of Interbull weighting factors for the Finnish test day model . Proc Interbull Tech Workshop, Verden, Germany 22-23.10.2000. *Interbull Bulletin 26*, 78-79.

Strandén, I. & Vuori, K. 2006. RelaX2: pedigree analysis program. *Proc. 8th WCGALP,* Belo Horizonte, Brazil.

Strandén, I. & Mäntysaari, E.A. 2010. A recipe for multiple trait deregression. Proc Interbull meeting, 31.5 – 4.6, 2010. Riga, Latvia. *Interbull Bulletin 42,* 21-24.

Tsuruta, S., Misztal, I., Aguilar, I. & Lawlor, T.J. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci. 94,* 4198-4204.

VanRaden, P.M. 2008. Efficient methods to compute genomic evaluations. *J. Dairy Sci. 91,* 4414-4423.

Table 1. Interbull GEBV test results for milk, protein and fat single-step evaluations for NAV RDC bulls. The PA is animal model parent average, GEBV $AM_{parameters}$ is GEBVs using the parameters from deregression, and GEBV $AM_{estimated}$ is GEBVs using the variances estimated from DRP data.

	Milk			Protein			Fat		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
PA	3.29	0.70	0.22	1.22	0.89	0.25	2.17	0.80	0.28
GEBV $AM_{parameters}$	3.80	0.72	0.35	4.23	0.81	0.38	3.29	0.79	0.45
GEBV $AM_{estimated}$	3.90	0.71	0.35	5.01	0.76	0.36	2.17	0.80	0.45

Table 2. Interbull GEBV test results for milk, protein and fat single-step evaluations for NAV RDC bulls. The PA is sire model parent average, DGV is direct genomic values from 2-step fit, GEBV SM is GEBVs using single step genomic model with sire deregressed proofs (Koivula *et al.*, 2011).

	Milk[1]			Protein[1]			Fat[1]		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
PA	3.28	0.73	0.19	4.26	0.77	0.20	2.34	0.83	0.23
DGV	3.15	0.76	0.30	4.51	0.77	0.31	2.23	0.85	0.40
GEBV SM	3.67	0.69	0.32	4.70	0.74	0.35	2.69	0.80	0.44

[1] heritabilities were h^2_{milk}=0.39, $h^2_{protein}$=0.39, and h^2_{fat}=0.36.

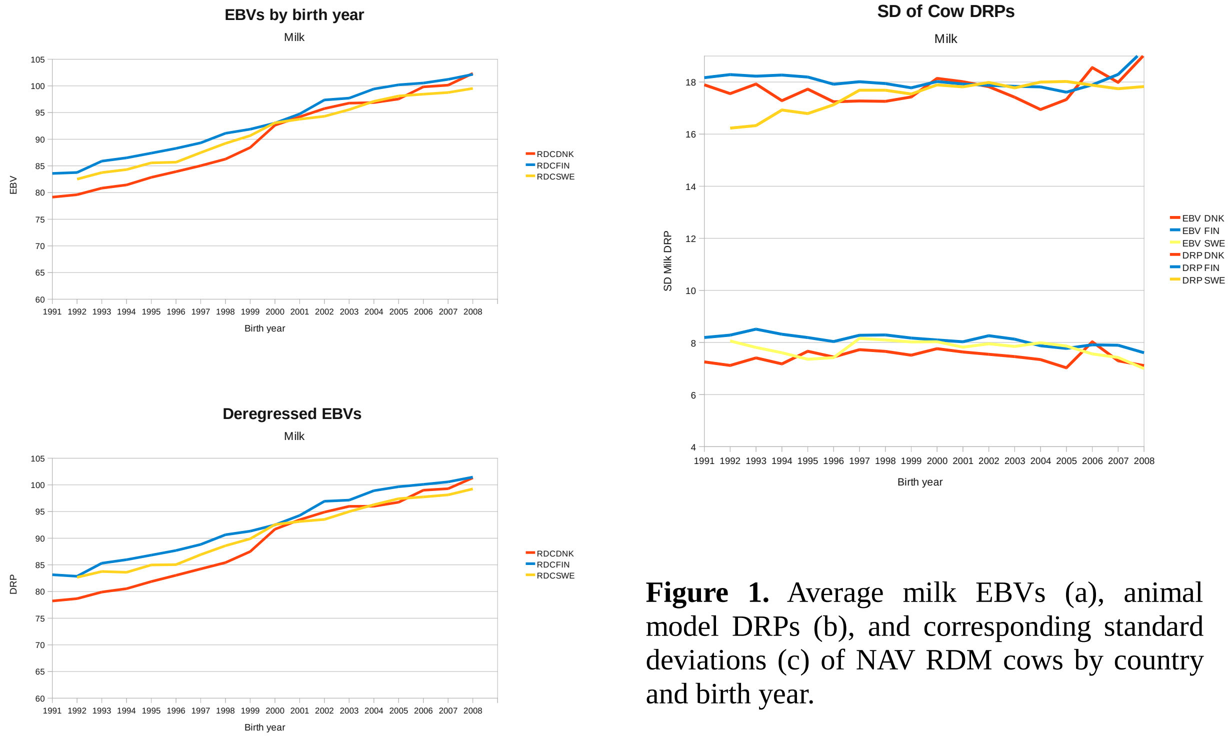

Figure 1. Average milk EBVs (a), animal model DRPs (b), and corresponding standard deviations (c) of NAV RDM cows by country and birth year.

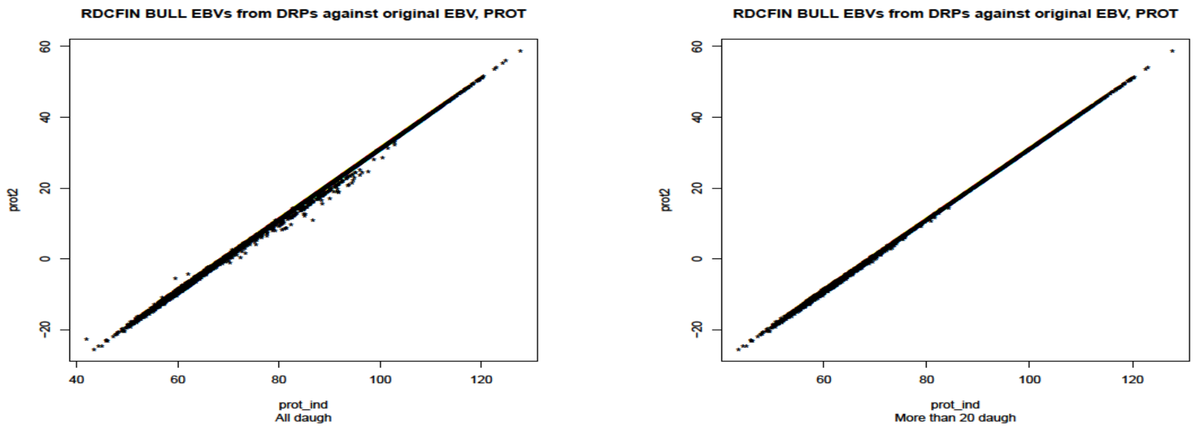

Figure 2. Protein EBVs of the Finnish registered RDC bulls recalculated from the animal model DRPs plotted against original EBVs. Top (a) figure with all bulls, bottom (b) figure with bulls having more than 20 daughters.