How ssGBLUP became suitable for national dairy cattle evaluations

I. Misztal[1*] D. Lourenco[1], S. Tsuruta[1], I. Aguilar[2], Y. Masuda[3], M. Bermann[1], A. Cesarani[1], A. Legarra[4]

[1] University of Georgia, Department of Animal and Dairy Science, 30602, Athens, GA, USA; [2] Instituto Nacional de Investigación Agropecuaria, 11500, Montevideo, Uruguay; [3] Rakuno Gakuen University, 069-8501, Ebetsu, Hokkaido, Japan; [4] Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, UMR1388 GenPhySE, 31326, Castanet Tolosan, France; [*]ignacy@uga.edu

Abstract

National dairy evaluations use mostly multistep methods, where a nongenomic BLUP evaluation is followed by extraction of pseudo-phenotypes, genomic analyses, and merging nongenomic and genomic evaluations. With genomic preselection and many females genotyped, pseudo-phenotypes are biased, and it is hard to accommodate female genotypes fully. Such problems were successfully solved in other species by ssGBLUP, but large biases and expensive computing were experienced in dairy. This paper documents several improvements that made ssGBLUP feasible for national dairy evaluations. The pedigree relationships account for inbreeding. The formulas for unknown parent groups involve relationships for genotyped animals, and the number of groups is small. Pedigrees and phenotypes are truncated. Large data is accommodated by the APY algorithm. Multibreed evaluation is possible with judicious choice of APY core animals. In tests with the US national data and up to 3.5 M genotypes, analyses by ssGBLUP showed minimal bias and superior stability of GEBV.

Introduction

Dairy genomic evaluations at the national level use mostly multistep methods (VanRaden et al., 2009), where a nongenomic BLUP evaluation is followed by extraction of pseudo-phenotypes for genotyped animals, genomic evaluation for genotyped animals, and an index to merge genomic and nongenomic values. Accurate creation of the index is difficult, and many countries use an approximate index, e.g., assuming 20% of polygenic and 80% of genomic component.

Over time as animals were genomically selected, BLUP became biased by pre-selection (Patry and Ducrocq, 2011; Masuda et al., 2018). Subsequently, pseudo-phenotypes are biased and require ad-hoc corrections. Accommodating genotypes of females without double counting is difficult, potentially creating problems as now most genotyped animals are females. One alternative to the multistep method is ssGBLUP (Aguilar et al., 2010; Misztal et al., 2020) that is used in almost all species except dairy. It automatically accounts for genomic preselection, avoids calculating pseudo-phenotypes, and accommodates male and female genotypes without double counting. In dairy, ssGBLUP was applied as early as 2010 by Aguilar et al. (2010) and more recently by Masuda et al. (2018) and Bradford et al. (2019); however, the evaluations were biased in all the cases. Another problem was a relatively large computing time due to slow iteration. Recently, two studies looked at an evaluation using Holsteins only (Cesarani et al., 2021) and multibreed using 5 breeds (Cesarani et al., 2022), and evaluations were nearly unbiased in both cases. An additional study compared evaluations in the last study with the official CDCB evaluations (Mota et al., 2022). That study found that the reliability of ssGBLUP was up to 0.07 higher, and stability was up by 0.12 higher. Subsequently, the current implementation of ssGBLUP seems beneficial for the routine evaluation in dairy cattle. The purpose of this paper is to describe steps that made ssGBLUP unbiased, accurate, and computationally efficient.

Materials & Methods

Data. Recently, several ssGBLUP evaluations for milk, fat, and protein were run with the CDCB data, only for Holsteins (Cesarani et al., 2021) or for five dairy breeds altogether that included Holstein, Jersey, Ayrshire, Brown Swiss, and Guernsey animals (Cesarani et al., 2022). Genotyped animals used in the analyses ranged from 800k to 3.9 million. The developments and improvements required to make ssGBLUP a suitable method for dairy evaluations are described below.

Compatibility between pedigree and genomic relationships. To avoid biases, pedigree and genomic relationships need to be compatible. Standard scaling is by a linear transformation of the genomic relationship matrix (**G**) for equal mean with the pedigree relationship matrix (**A**) (Vitezica et al., 2011). While **A** is affected by the completeness of pedigree, **G** is not. Therefore, an important part of scaling is accounting for inbreeding in the pedigree relationships. If pedigrees are missing, the inbreeding of animals with missing parents is calculated as 0, so accounting for nonzero inbreeding of missing parents (VanRaden, 1992) could provide additional compatibility.

Unknown Parent Groups (UPG). Under selection, the merit of missing parents over time needs to be accounted for. Early studies in ssGBLUP accounted for UPG only in **A**, resulting in slow convergence and biases. Later studies found that a reasonable option is also considering UPG for pedigree relationships for genotyped animals ($\mathbf{A}_{22}$). Many BLUP evaluations use 100s of UPG. With UPG in $\mathbf{A}_{22}$, there is not enough information to account for many groups, and considerably reducing the number of groups was a good option to reduce bias.

Interaction among pedigree, data truncation, and UPG. In general, the additive information decays fast with subsequent generations, and old data can be discarded without losses in accuracy for new animals but with gains in computing time. Including more than 2 generations of data with additional 2 generations of pedigree did not improve the accuracy for young animals and sometimes even lowered it (Lourenco et al., 2014; Cesarani et al., 2021). The broiler industry uses at most 3 generations of data despite many more available. Potential reasons for lower accuracy with older data are changing definitions of traits, model adjustments, and poor convergence for the solutions for young animals. Additionally, age adjustments vary over time as cows are indirectly selected for early calving. Truncation of old pedigrees eliminates the effect of missing pedigrees prior to truncation and reduces the number of needed UPG. If the number of generations is small, accounting for inbreeding is less or not critical.

Metafounders (MF) in ssGBLUP. MF are generalizations of UPG with scaling and inbreeding plus covariances accommodated (Legarra et al., 2015). In regular scaling, **G** is scaled for compatibility with $\mathbf{A}_{22}$. With MF, **G** is created with 0.5 gene frequencies, and pedigree relationships are adjusted for compatibility with **G**. Variance and covariances across MF potentially account for inbreeding of unknown parents and relationships among unknown parents. While MF did very well in simulations, problems with parameter estimation are still under investigation in real populations.

Large number of genotypes. The dairy populations accumulated a large number of genotypes, and many algorithms exist that can support that number in single-step methods (Misztal, 2018). All of them exploit the fact that the rank of genomic computations (based on SNP or **G**) is limited by the number of SNP (usually 60k) and by the small dimensionality of **G** in populations with small effective population size. This dimensionality is about 15k in dairy cattle (i.e., 15k core animals). With the APY algorithm (Misztal, 2016), which relies on the limited dimensionality of **G**, calculations have linear computing and storage costs past the number of core animals. An advantage of APY is the low cost and easy application of UPG. A disadvantage is some fluctuations of GEBV with the choice of core animals, which can be handled by keeping the core constant. An important part of the evaluation is pruning genotypes. While the number of genotyped animals may be very high, those without progeny or phenotypes can be eliminated from ssGBLUP evaluations, reducing computing costs. Predictions for omitted animals by indirect predictions can be as accurate with large data sets as when they are included in the model (Tsuruta et al., 2021). Another factor influencing computations is the blending of **G**, used to avoid the singularity of **G**. Blending with $\mathbf{A}_{22}$ was costly; however, computing coefficients separately for core and noncore animals is highly efficient. Blending with $0.01\mathbf{I}$ is also efficient without affecting accuracy.

Other issues. Many other issues are remaining in the dairy ssGBLUP evaluations that are not discussed here but have been addressed elsewhere. They include using external information from Interbull (Guarini et al., 2019), exploiting putative QTN (Misztal et al., 2020), using purebred and crossbred information combined (Misztal et al., 2022), and approximating accuracies of GEBV (Bermann et al., 2022).

Results

Holsteins (Cesarani et al., 2021). Phenotypes were truncated for cows recorded before 1980, 1990, or 2000, and pedigrees were truncated to 2 or 3 generations beyond phenotyped cows. Two UPG formulations were used, the QP transformation for **A** (UPG1) or **A** and $\mathbf{A}_{22}$ (UPG2). With forward prediction, reliabilities were calculated for bulls and predictivity for cows. Only genotypes for contributing animals were used (about 800k). With UPG2, reliabilities and accuracies were the same with all truncations. Inflation was minimal as the b_1 parameter was close to 1.0. With simple UPG, reliabilities were smaller and varied with truncation.

Five breed analyses (Cesarani et al., 2022). Holstein, Jersey, Ayrshire, Brown Swiss, and Guernsey data were analyzed separately and altogether in a model with UPG2, truncation of phenotypes recorded before 2000, and 3 generations beyond phenotypes. The total number of genotyped animals was 3.9 million. When the multibreed evaluation used 15k randomly selected core animals in the APY algorithm, the reliabilities were lower for the smaller breeds than in the single-breed analyses. When core animals were selected to include 5k animals of each smaller breed and 15k of each Holstein and Jersey, the reliabilities were close to those from single breeds, and they were higher for Jerseys. One explanation is that the Jersey population included Holsteins in the past. In all cases except for the smallest breeds, the inflation was low (i.e., b_1 close to 1.0).

Official comparisons. Evaluations for Holstein bulls from the previous analyses were compared with official multistep evaluations (MS) by CDCB (Mota et al., 2022); the latter were computed using different weights for SNP in a nonlinear approach, whereas ssGBLUP assumed equal weights for SNP. For milk, R^2 based on DYD for raw MS was 0.07 lower than for ssGBLUP. After age adjustments, this difference reduced to 0.01. Stability of evaluations was calculated by correlations of GEBV derived with 2017 and 2021 data. The stability for MS was 0.86 (0.89 with correction for age). The stability for ssGBLUP was 0.90 (0.92 after age correction). Regression coefficients were close to 1.0 for both ssGBLUP and MS predictions.

Discussion

The results for reliability and stability indicate a dilemma on how to judge an evaluation system. If DYD is derived from BLUP, it is biased even for bulls with many daughters. Also, GEBV by unadjusted MS are biased. Subsequently, it is not clear whether adjustments for age compensate GEBV, DYD, or both. In terms of stability, ssGBLUP was ahead of MS. After many developments as described, ssGBLUP is finally ready for implementation at the national level. This method provides evaluations that are possibly more accurate than the current one, does not require adjustments, and is simpler to run. Further improvements of accuracy can be expected with the inclusion of external Interbull data and possibly weighing SNP differently.

References

Aguilar I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, et al. (2010) J. Dairy Sci. 93:743-752. doi:10.3168/jds.2009-2730

Bermann M., D. Lourenco, and I. Misztal. (2022) J. Anim. Sci. skab353. doi:10.1093/jas/skab353

Bradford H. L., Y. Masuda, J. B. Cole et al. (2019) J. Dairy Sci. 102:2308-2318. doi:10.3168/jds.2018-15419

Cesarani A., Y. Masuda, S. Tsuruta, et al. (2021) J. Dairy Sci. 104:5843-5853. doi:10.3168/jds.2020-19789

Cesarani A., Y. Masuda, S. Tsuruta, et al. (2022) J. Dairy Sci. (Under review)

Guarini A.R., D.A.L. Lourenco, L.F. Brito, et al. (2019) J. Dairy Sci. 102:8175-8183. doi:10.3168/jds.2018-15819

Lourenco D., I. Misztal, S. Tsuruta, et al. (2014) J. Dairy Sci. 97:3930-3942. doi:10.3168/jds.2013-7769

Legarra A., O. F. Christensen, Z. G. Vitezica, et al. (2015) Genetics 200:455-468. doi:10.1534/genetics.115.177014

Mota R.R, A. Cesarani, and P.M. VanRaden. Proc. of the 12nd WCGALP, Rotterdam, The Netherlands.

Masuda,Y., P. M. VanRaden, I. Misztal, et al. (2018). J. Dairy Sci. doi:10.3168/jds.2017-13310

Misztal I. (2018) 11th WCGALP, Auckland, New Zealand.

Misztal I., D. Lourenco, and A. Legarra (2020) J. Animal Sci. 98:1-14. doi:10.1093/jas/skaa101

Misztal I. (2016) Genetics 202:401–409. doi:10.1534/genetics.115.182089

Misztal I., Y. Stein, and D.A.L. Lourenco (2022) J. Dairy Sci. Comm. (Accepted)

Patry C., and V. Ducrocq. (2011) Genet. Sel. Evol. 43:30. doi:10.1186/1297-9686-43-30

Tsuruta S., D. Lourenco, Y. Masuda, et al. (2021) J. Dairy Sci. Commun. 2:356-360. doi: 10.3168/jdsc.2021-0097

VanRaden P. M. (1992) J. Dairy Sci.75:3136–3144. doi:10.3168/jds.S0022-0302(92)78077-1

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, et al. (2009) J. Dairy Sci. 92:16-24. doi:10.3168/jds.2008-1514

Vitezica Z., I. Aguilar, I. Misztal, and A. Legarra (2011) Genet. Res. Camb. 93:357-366. doi:10.1017/S001667231100022X