Avoiding Bias From Genomic Pre-Selection in Converting Daughter Information Across Countries

P.M. VanRaden
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA

Abstract

Methods to include both foreign and genomic information in single-step or multi-step evaluations were developed and compared using the U.S. national Jersey database. Breeders have exchanged and converted genetic evaluations of bulls across countries for decades, but traditional evaluations may become biased by pre-selection on genotype. When foreign and genomic data were added to the equations, daughter yield deviations computed from only domestic daughter records were very stable. Those could be exchanged internationally, thereby avoiding the difficulty of deregressing genomic evaluations. A final step in the multi-step method simply inserted the genomic evaluations and held them constant during iteration instead of adjusting the data vector and equations. For genotyped young bulls, multi-step evaluations were correlated by .966 to single-step evaluations computed with an algorithm that did not require inverting the genomic relationship matrix. Accuracy was similar but regressions were closer to expectation for the single-step evaluations.

Key words: single-step genomic evaluation, foreign data, selection bias

Introduction

Exchange of phenotypic information using multi-trait across-country evaluation (MACE) has allowed foreign bulls to be easily included in genomic reference populations when their genotypes are exchanged. Traditional domestic evaluations are the first step, and MACE evaluations of foreign data are the second step in multi-step genomic evaluation. Those methods combine pedigrees and phenotypes first, and then information from genotypes is added later.

Traditional models that do not account for genomic selection may become severely biased (Vitezica *et al.*, 2010; Patry and Ducrocq, 2011b). Traditional MACE was not affected by genomic pre-selection before 2011; however, bulls born in 2008, sampled in 2009, and with daughter records in 2012 were pre-selected on genotype and may require new exchange methods. Genotypes as an additional data source can greatly improve accuracy and timeliness of selection, but optimum methods and algorithms to solve large sets of equations and include foreign data are not yet fully developed. National evaluations that combine all available data sources simultaneously can be more accurate but also more difficult to set up and solve.

The single-step method can be applied to large national data sets (Aguilar *et al.*, 2010), but computations quickly become a limiting factor as numbers of genotyped animals increase. Multi-trait evaluations were affordable for a type data set with 6 million phenotyped and 16,900 genotyped animals (Tsuruta *et al.*, 2011). However, about 30 million animals have phenotypes in U.S. yield evaluations, and over 150,000 now have genotypes, with that number expected to double again this year. Matrix inversion costs are cubic with number of genotyped animals and already are not feasible.

A mathematically equivalent but less costly approach was proposed by Legarra *et al.* (2011). Their algorithm appends extra equations that include the genomic relationship matrix instead of its inverse and the pedigree relationship matrix for genotyped animals instead of its inverse to the mixed model equations. Computation is linear rather than cubic with number of genotyped animals. Although the math seems valid, the equations are not positive definite, and the iterative strategy has not been applied to real data sets yet.

An alternative approach proposed by Mäntysaari and Strandén (2010) and tested using Dutch national data by Stoop *et al.* (2011) includes genomic information as a separate correlated trait. A second alternative is to include the multi-step genomic EBV (GEBV) minus EBV difference as additional data for the same trait (Patry and Ducrocq, 2011a). Those alternative approaches are not as appealing in theory because genomic calculations are done separately and require traditional input data that may become biased by pre-selection, but they may be more practical than single-step algorithms.

Foreign information has been directly included in some traditional national evaluations using pseudo-records for daughters of foreign bulls in the mixed model (Bonaiti and Boichard, 1995; Pedersen *et al.*, 1999). Methods to include foreign data in single-step genomic evaluation had not been developed yet. Incorporation of genomic and foreign information in U.S. genetic evaluations required complete revision of computer software to allow more multi-trait processing. Some U.S. trait evaluations are single trait, some are exact multi-trait, and others (such as productive life) use approximate multi-trait post-processing. A unified multi-trait analysis of all traits is still not possible because of the use of several different models and the mixture of normal and non-normal traits.

This report outlines methods for maintaining unbiased exchange of phenotypic information across countries. That problem is separate from genomic MACE (GMACE; VanRaden and Sullivan, 2010) or simple GMACE (Sullivan, 2011). The goal of GMACE methods is to convert national GEBVs from one country to another. The goal of the current research is to ensure that unbiased phenotypic information from foreign EBVs continues to be available as in traditional MACE.

Methods

Daughter yield deviation (DYD) accounts for merit of mates (EBV_{mate}) and herdmates:

$$DYD = \sum q(w)[YD - 0.5(EBV_{mate})]/ \sum q(w),$$

where q is 2 if mate is known or 4/3 if mate is unknown, w is a weight for number of records of each daughter, and YD is yield deviation.

Genomic DYD (DYD_g) can account for genomic merit of mates ($GEBV_{mate}$) and herdmates in a single-step evaluation simply by substituting terms to obtain:

$$DYD_g = \sum q(w)[YD_g - 0.5(GEBV_{mate})]/\sum q(w),$$

where each daughter's genomic yield deviation (YD_g) is defined as the weighted sum of a cow's records adjusted for environmental effects, the same as traditional YD, except that the environmental effects are solved together with genomic information to prevent bias from pre-selection of bulls. Some bias may still occur in DYD_g if the bull's daughters are also pre-selected and only those with better genomic merit receive phenotypes.

Foreign information was included using one record weighted by daughter equivalents for each bull that had foreign daughters instead of one pseudo-record for each foreign daughter. The method of Bonaiti and Boichard (1995) was also modified for multi-trait models by pre-multiplying the vector containing DYD for each trait by the inverse of the genetic covariance matrix among traits. The foreign DYD ($DYD_{foreign}$) was estimated from the MACE EBV using the simple one bull at a time method:

$$DYD_{foreign} = PA_{IB} + (EBV - PA_{IB})/REL_{IB},$$

where PA_{IB} and REL_{IB} are parent average and reliability from Interbull. For bulls with both domestic and foreign daughters, the foreign portion of DYD was obtained by replacing PA_{IB} in the formula above with domestic EBV and computing REL_{IB} using MACE minus domestic daughter equivalents. Matrix de-regression might be better.

Genomic information was included using two different methods. The first method was single-step GEBV computed using the equations of Legarra *et al.* (2011). The second method computed multi-step evaluations as in VanRaden *et al.* (2009), and then the GEBVs were inserted into animal model equations and held constant while solving for all other effects. That approach differs from earlier

studies such as Patry and Ducrocq (2011a) or Mäntysaari and Strandén (2010) because the data vector and mixed model equations are not adjusted but EBVs of all other animals are adjusted by pedigree relationships with animals with GEBVs.

The U.S. national database from December 2011, which contained 4.4 million lactation records for milk yield of 1.5 million Jerseys and genotypes for 5,364 males and 11,488 females, was used to test the methods and algorithms. Foreign DYDs from 7,072 bulls were either excluded or included along with national phenotypes. Genomic information was excluded or included by either single-step or multi-step methods in equations that also included foreign DYDs. The complete pedigree file of 4.1 million animals including old and young, domestic and foreign was used in all evaluations. Crossbred daughters were excluded from this study but are included in official all-breed evaluations. To test accuracy, phenotypes were truncated in August 2007, and the same methods were applied to predict current data.

Results and Discussion

Correlations between DYD_g and DYD in Table 1 were very high for U.S. bulls (>0.9993) regardless of inclusion of foreign or genomic information in the system of equations. Those correlations could decrease with pre-selection in the future. The summation across daughters did not include the pseudo-record for foreign daughters so that only domestic daughter information was included in the bull's DYD_g or DYD. Exchanging those in MACE may be a simple way to account for genomic pre-selection in national evaluations and continue to provide unbiased traditional information to foreign partners. However, that approach is not as simple if the national evaluation includes additional genetic effects such as separate parities that are not exchanged in MACE.

Correlations between GEBVs for young U.S. bulls in Table 2 were fairly high (0.966) between single-step and multi-step methods as compared with correlations between GEBV and PA (0.853 to 0.869). The PAs with and without foreign data for U.S. young bulls were highly correlated (0.997) because most had U.S. sires and because foreign dam EBVs were not included in the study. The MACE EBVs of foreign sires were correlated by 0.77 with their EBVs using only national data but increased to 0.995 after including foreign sire DYD, indicating that the simple method was successful.

Table 1. Correlations among DYDs computed with or without genomic and foreign data in the model.

Foreign data		No	Yes	Yes	Yes
	Genomic data	No	No	Single-step	Multi-step
No	No	1.0	0.9998	0.9993	0.9997
Yes	No		1.0	0.9993	0.9996
Yes	Single-step			1.0	0.9997
Yes	Multi-step				1.0

Table 2. Correlations among young bull PAs excluding or including foreign data and single-step or multi-step GEBVs.

Foreign data		No	Yes	Yes	Yes
	Genomic data	No	No	Single-step	Multi-step
No	No	1.0	0.997	0.868	0.856
Yes	No		1.0	0.869	0.853
Yes	Single-step			1.0	0.966
Yes	Multi-step				1.0

Predictions from August 2007 data had squared correlations with future DYD of 0.436 for PA, 0.520 for multi-step, and 0.520 for single-step evaluations. Corresponding regressions were 0.73, 0.75, and 0.85, all lower than the expected regression of 0.93. The 2007 truncated reference population included 2,029 bulls and 987 cows, whereas the current reference population included 2,561 bulls and 5,620 cows.

Preliminary results for Holstein data revealed much slower convergence of the single-step algorithm than for Jersey data, and the algorithm diverged for multi-trait equations with more than four traits. Second-order Jacobi iteration with block diagonal solution was used in the study, but pre-conditioned conjugate

gradient iteration could improve results as recommended by Tsuruta *et al.* (2001). Convergence of the equations of Legarra *et al.* (2011) is not guaranteed with just any algorithm; however, from theory (Broyden, 1964), a scheme with (block) successive under-relaxation does ensure convergence (A. Legarra and V. Ducrocq, INRA, France, personal communication). Further algorithm development and testing are needed because single-step evaluations look promising when they converge.

For future international exchange, countries could compute national single-step genomic evaluations, possibly including foreign data, and provide the DYD_g from domestic daughters for conversion using MACE. Then, new foreign DYD_g free from selection bias could be incorporated into the national single-step equations, replacing any foreign information from the previous evaluation. To obtain GEBVs for foreign bulls without genotypes included in domestic data, separate exchange methods (such as GMACE or simple GMACE) will continue to be needed.

Very old bulls are not included in traditional MACE exchange, but many are now being genotyped or sequenced because current animals have many copies of their genes. The North American database includes genotypes for 479 bulls born before 1985 that have been traded with foreign partners but are difficult to include in reference populations because of being excluded from MACE. Inclusion of those bulls could improve genomic reliability slightly.

Female phenotypic information is used more fully in some countries than in others. The PAs may include or exclude domestic or foreign dam EBVs, but most multi-step evaluations have not included females in the reference population. Exchange of foreign cow and dam EBVs continues to require much effort, and partners must be sure to convert and include EBVs rather than GEBVs for genotyped females to avoid double counting the genomic information.

Conclusions

Genotype exchange partners need unbiased phenotypic information from foreign reference animals to compute unbiased genomic evaluations, but MACE inputs may become biased by genomic pre-selection. Simultaneous equations can include phenotype, genotype, pedigree, and foreign information together. In such systems, DYD_g can account for pre-selection of bulls and genomic merit of herdmates when summarizing daughter information, but may still contain bias if daughters are pre-selected on genotype before being phenotyped. Exchange of DYD_g across countries could eliminate the need to partition genomic from phenotypic information at Interbull and the need to deregress and reregress evaluations. Methods were tested on U.S. Jerseys, but application to Holsteins will require revision of the algorithm to speed convergence.

Acknowledgements

The author thanks Andres Legarra, Ignacy Misztal, Tom Lawlor, Ignacio Aguilar, Shogo Tsuruta, Vincent Ducrocq, and George Wiggans for helpful comments and excellent advice on computation, even though unfortunately all their advice has not been followed yet, and for sharing subroutines used in the single-step algorithm. The U.S. industry organizations represented by the Council on Dairy Cattle Breeding provided the data for the study.

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.

Broyden, C.G. 1964. On convergence criteria for the method of successive over-relaxation. *Math. Comp. 18*, 136–141.

Bonaiti, B. & Boichard, D. 1995. Accounting for foreign information in genetic evaluation. *Interbull Bull. 11,* 4 pp.

Legarra, A., Misztal, I. & Aguilar, I. 2011. The single step: Genomic evaluation for all. *Book of Abstr. of 62nd Annu. Mtg. of Euro. Fed. of Anim. Sci. No. 17,* 1. Wageningen Academic Publishers, The Netherlands.

Mäntysaari, E.A. & Strandén, I. 2010. Use of bivariate EBV-DGV model to combine genomic and conventional breeding value evaluations. *Proc. 9th World Congr. Genet. Applied to Livest. Prod.,* Comm. 353.

Patry, C. & Ducrocq, V. 2011a. Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genet. Sel. Evol. 43,* 30.

Patry, C. & Ducrocq, V. 2011b. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci. 94,* 1011–1020.

Pedersen, G.A., Pedersen, J., Nielsen, U.S. & Madsen, P. 1999. Experiences of blending foreign information in the national genetic evaluation. *Interbull Bulletin 22,* 61–65.

Stoop, W.M., Eding, H., van Pelt, M.L., & de Jong, G. 2011. Combining genomic and conventional data in the Dutch national evaluation. *Interbull Bulletin 44,* 169-172.

Sullivan, P.G. 2011. Accounting for residual correlations among regional genomic predictions via GMACE. *Interbull Bull. 43,* 19-24.

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.

Tsuruta, S., Misztal, I. & Strandén, I. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci. 79,* 1166–1172.

VanRaden, P.M. & Sullivan, P. 2010. International genomic evaluation methods for dairy cattle. *Genet. Sel. Evol. 42,* 7.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92,* 16–24.

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genet. Res., Camb. 93,* 357–366.