Robust GMACE for Young Bulls - Methodology

***P.G. Sullivan*[a] *and J.H. Jakobsen*[b]**
[a]*Canadian Diary Network, Guelph, ON, Canada,*
[b]*Interbull Center, SLU, Uppsala, Sweden*

Abstract

Methods presented previously to combine GEBV of young bulls and MACE solutions of ancestors were reviewed. Variances required for GMACE could be assumed equal to the variances in regular MACE, or estimated from GMACE input data. Equations to estimate "genomic" variance were partitioned to explain extreme estimates that have been observed. Variance estimation was subsequently improved and constraints were applied to avoid extreme variances in GMACE applications. Subtracting the average difference between national GEBV and MACE parent average forced a null average for Mendelian Sampling estimates and removed inconsistencies among population scales. This adjustment reduced or eliminated the majority of extreme genomic variance estimates. The small number of remaining extremes were for traits with unusually low reliabilities of national GEBV. Nearly all other estimates of genetic standard deviation (SD) were within the range 0.80-1.20 times the SD used for MACE. Estimates outside this range were truncated to the edges of the range. RMSE of local GEBV predictions, based on GMACE of data that included GEBV from only foreign countries, were reduced by these constraints on genomic variance estimates. The use of robust variance estimates also reduced the bias of top young bull predictions, especially for traits with the largest biases. Relative to the use of MACE variances, GMACE with robust genomic variances gave a slightly higher but similarly low maximum bias for SCS (20% versus 18%) and for all other traits the maximum bias was reduced, from 22% to 10% for protein yield, from 46% to 16% for stature, from 61% to 44% for longevity, and from 28% to 27% for fertility.

Key words: genomics, international evaluation, MACE, GMACE, robust

Introduction

Methods to compute international genomic evaluations of young bulls were described and compared by Sullivan *et al.* (2011). Estimation of genomic variances was recommended but with a caution that poor estimates of variance for some traits and countries could cause problems.

The purpose of this study was to investigate possible reasons for poor variance estimates, to improve the variance estimation methods if possible, and to consider ways to minimize the risk of poor GMACE results if input data are less than ideal for some countries or traits.

Data

The data and edits for the present study are described in detail by Jakobsen and Sullivan (2012). In summary, there were five traits included: protein (pro), stature (sta), somatic cells (scs), direct longevity (dlo) and female fertility (cow conception trait #1; cc1). August 2011 national GEBV data from eleven populations (CAN, DEU, DFS, FRA, NLD, POL, USA, CHE, CHR, ITA and JPN), and EBV data from all countries participating in the August 2011 MACE service of Interbull were used for the present study. The total numbers of national GEBV on young genotyped bulls without daughter data, across all populations, were: 57902 for pro, 47285 for sta, 53820 for scs, 54663 for dlo, and 44395 for cc1.

Methods

Methods for international genomic evaluation have evolved following several presentations, international discussions and Interbull pilot studies, of which the present is GMACE pilot study number 3. Many different acronyms have been used to describe approaches that are only minor variations on the GMACE approach first described by Van Raden and Sullivan (2010). A new set of acronyms is proposed in the present paper (Table 1) to more clearly reflect

the complete evolution of methods and specific changes that were made. Method GM_all, described by Van Raden and Sullivan (2010) allows all available GEBV from all countries and for all bulls to be used as input data. Where GEBV are not available, EBV data are used. The residual correlation matrix in GM_all was modified by Sullivan and Van Raden (2010), but we do not need to introduce a new acronym for this minor change.

The SGMACE method (Zumbach *et al.*, 2011) was an attempt to allow GEBV as input data in a regular MACE application, by restricting to only 1 GEBV per bull (old or young), and otherwise use all available EBV on any other country scale. Method GM_yng uses the same methodology as GM_all but with a restriction on GEBV input data to use only the GEBV of young bulls with no daughters and EBVs of proven bulls.

Method GM_ms signifies a more significant change in methodology relative to GM_all and GM_yng. With GM_ms, the methods described by Van Raden and Sullivan (2010) are still applied, but to an individual bull's equations rather than simultaneously for all bulls in the population. Reasons for this change were discussed in detail by Sullivan (2011). GM_ms(v) is exactly the same method as GM_ms but with a genomic variance estimation step included, rather than assuming GEBV-based genomic variances and EBV-based MACE variances are equivalent. Method rGM_ms(v) additionally adds the robust changes described in the present paper, which are to avoid the use of poor (extreme) estimates of genomic variance.

GMACE of Mendelian Sampling (GM_ms)

For each animal, a set of Mendelian Sampling (MS) mixed-model equations are set up (**Lu=r**), with **u** representing the vector of MS values for all countries. For the available GEBV, MS are calculated as [**GEBV – MACE PA**], and otherwise as [**MACE – MACE PA**], where **PA** denotes parent average. Given the left-hand-sides (**L=[D+G^{-1}]**) and corresponding solutions (**u**), the right-hand-sides are derived as [**r=Lu**]. Noting that [**r=Dy**], pseudo observations are thus derived as [**y=D^{-1}r**]. Matrix D includes traditional EDC and genomic GEDC data for the bull, and in this case EDC=0 for the young bulls if we ignore contributions of the bull's dam to a bull's EDC. We note here that ignoring the dam's own data contribution can lead to underestimates of reliability from GMACE in some cases, but that EBV contributions from dams are routinely excluded from international evaluations due to potential biases in elite cow EBVs.

Finally, matrix D is replaced with matrix E* as described in Sullivan and VanRaden (2010), and the modified equations are solved to derive international MS estimates for all countries, to which we add back the MACE PA that were originally subtracted.

Genomic Variances for GM_ms(v)

Countries attempt to generate GEBV that are directly comparable to EBV of non-genotyped bulls. However, different methods and assumptions for genomic prediction, and for deriving GEBV as a combination of direct genomic values (DGV) and EBV, can lead to different GEBV variances in each country. Additionally, the relative consistency between GEBV and MACE PA in each country can affect the variance of genomic MS deviations.

To estimate and account for these differences, genomic variances can be estimated via REML as is routinely done for genetic or sire variances in regular MACE (Sullivan, 1999). Inputs required are MS estimates and prediction error variance (PEV) of MS. The latter term is a quadratic function of the PEV matrix for animal, sire and dam, which can be approximated as the corresponding matrix from MACE reliability approximation within the software (i.e. the inverted LHS after absorbing all other relatives), with rows and columns rescaled by the relative change in PEV due to additional GEDC. The relative change due to GEDC is the ratio of animal PEV that corresponds with total effective records (MACE+GEDC) divided by animal PEV that corresponds with effective records from only MACE.

To investigate extreme genomic variances, as estimated above, it is useful to re-write the estimation equation in the equivalent form:

$$V(g) = 2 * ave[\hat{M}^2] / ave[Rel(\hat{M})]$$

where V(g) is genomic variance, ave[] denotes an average, Rel denotes reliability, $\hat{M}$ is a genomic estimate of Mendelian sampling as a deviation from MACE parent average, and 2 is the ratio of genetic to MS variance when both parents are known. High V(g) can be expected if numerator values are high or denominator values are low.

Cross Validation

Young bulls with national GEBV from multiple countries were used to cross-validate the international genomic evaluation methods. Assessing one country scale at a time, all local GEBV were deleted and then predicted with GMACE from the remaining data, which included foreign-scale GEBV of young bulls plus EBV from all scales for proven bulls (e.g. sires of the young bulls). Root mean squared errors (RMSE) of prediction were computed, and also an estimate of bias at the upper end (top) of the distribution of young bulls. A top bull was taken as having a GMACE result (*z*) of 3 standard deviations above the mean on a given country scale. The cross-validation prediction equation, from regressing local GEBV on the foreign-data predictions, was then used to derive an expected local GEBV (*w*) for such a top bull. Top bull bias was computed as the percentage error in the prediction:

$$\text{Top Bull Bias} = (100\% * (z-w)/w)$$

Results and Discussion

As should be expected, estimates of genomic variance were generally close to the traditional genetic variances used in MACE. With the exception of 6 outlier estimates among the 45 in total across all traits and populations, genomic variance estimates based on GEBV of young bulls were on average almost exactly equal to the genetic variances from MACE based on EBV of proven bulls. The 6 outlier estimates were partitioned into numerator and denominator factors in Table 2. Three of the 6 outliers were due to an inflated numerator, and two were due to an unusually small denominator.

The inflated numerators could be due to selective genotyping, selective reporting of GEBV, or inconsistencies between MACE of selected parents and the national GEBV of young bulls. Because of the clear relationship between a non-zero average for MS and inflated genomic variance estimates, a simple improvement was to force an average of zero on the MS estimates prior to the MS-conversions among population scales. This adjustment ensures an average consistency between (adjusted) GEBV and MACE for all countries. Efforts should also be made to determine at the national level if there are issues of GEBV selection or problems of consistency with MACE, which can in some way be corrected.

The small denominators were due to unusually low reliabilities for the national GEBV. Although lower than the other populations in this study, there was little reason to believe the reliabilities were incorrect, as these were from relatively small reference populations in both cases. At lower levels of reliability, even small errors in the national reliability approximations will have much larger effects on genomic variance estimates, because of division by small fractional values. The approximation of PEV(MS) is also subject to greater errors when mixing MACE solutions of parents and young bull GEBV of low reliability, and it is more likely to have GEBV-MACE inconsistencies that are not handled perfectly with the variance approximation that is used. Options in this case are to exclude such data from GMACE, find ways to improve the approximations used for variance estimation or simply constrain the extreme variance estimates to more reasonable values. Although somewhat arbitrary, the latter option was used as a simple way to minimize potential problems from these poor variance estimates, while still allowing all interested countries to participate in the GMACE evaluation. Estimated genomic SD were truncated at a maximum difference of ± 20% relative to the genetic SD from MACE.

Results for RMSE (Table 3) and top bull bias (Table 4) are presented for three of the more recent GMACE methods described in Table 1. Differences among the methods were relatively small for RMSE of most traits and countries, but for Stature the robust method of the present study (rGM_ms(v)) was clearly

preferred. Advantages of rGM_ms(v) were more obvious when considering Top bull bias, which was generally closer to zero for many traits and countries, relative to the other GMACE methods. The RMSE results show that rankings among young bulls will be similarly accurate for all methods. The top bull bias results show that comparisons of young with proven bulls will likely be better with the rGM_ms(v) method.

The robust adjustments would be most useful for small populations, and it should be noted that cross-validations were by necessity limited to the larger populations. Results presented here likely understate the usefulness of these adjustments.

References

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci. 77,* 2671-2678.

Sullivan, P.G. 1999. REML estimation of heterogeneous sire (co)variances for MACE. *Interbull Bulletin 22,* 146-148.

Sullivan, P.G. 2011. Accounting for residual correlations among regional genomic predictions via GMACE. *Interbull Bulletin 43,* 15-20.

Sullivan, P.G., Zumbach, B., Dürr, J.W. & Jakobsen, J.H. 2011. International genomic evaluations for young bulls. *Interbull Bulletin 44,* 87-94.

Sullivan, P.G. & VanRaden, P.M. 2010. GMACE implementation. *Interbull Bulletin 41,* 3-7.

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

Zumbach, B. Jakobsen, J., Forabosco, F., Jorjani H. & Dürr, J. 2011. Data selection and pilot run on Simplified Genomic MACE (S-GMACE). *Interbull Bulletin 43,*11-18.

Table 1. International genetic and genomic evaluation terminology.

Method	Model	Genetic Variance Variable	Genomic Input Data (EBV are used otherwise)	Remarks
MACE	MACE	EBV	No GEBVs	Schaeffer, 1994
SGMACE	MACE	EBV	1 GEBV per bull, all bulls	Zumbach *et al*, 2011
GM_all	GMACE	EBV	All GEBV, all bulls	VanRaden and Sullivan, 2010
GM_yng	GMACE	EBV	All GEBV, all young bulls	Sullivan *et al*, 2011 (GMACE)
GM_ms	GMACE	EBV	All GEBV, 1 young bull	Sullivan *et al*, 2011 (MCNV)
GM_ms(v)	GMACE	GEBV	All GEBV, 1 young bull	Sullivan *et al*, 2011 (VCNV)
rGM_ms(v)	GMACE	GEBV	All GEBV, 1 young bull	Present paper, robust constraints on variances

Table 2. The 6 outliers and the range of all 39 other SDratios[z], and variables contributing to the numerator (Ave ($\hat{M}^2$)) and denominator (Rel($\hat{M}$)) of the variance estimation equation. Multiple values for the 39 non-outliers are means surrounded by the 95% confidence intervals (L95, mean, U95).

SDratio	Ave ($\hat{M}^2$)	Ave($\hat{M}$)	Rel($\hat{M}$)
3.14	2.56*	-0.87*	0.42
2.66	3.29*	-1.16*	0.53
1.90	0.69	-0.08	0.08*
1.83	0.68	0.07	0.43
1.57	0.56	0.20	0.16*
1.51	1.62*	0.43*	0.48
0.71 - 1.29	0.06, 0.56, 1.07	-0.24, 0.03, 0.29	0.28, 0.46, 0.65

[z]Genomic SD from GEBV of young bulls / Genetic SD from EBV of proven bulls

*Value is outside the 95% C.I. of non-outliers, which were SDratios in the range 0.71 to 1.29

Table 3. RMSE (*100) for three GMACE predictions (GM_yng, GM_ms, rGM_ms(v)).

Country	Protein	Stature	SCS	Longevity	Fertility
CAN	23, 24, 19	76, 84, 59	27, 27, 26	57, 57, 49	54, 54, 53
USA	27, 25, 24	n/a	29, 29, 29	51, 56, 53	n/a
NLD	24, 24, 25	31, 31, 29	36, 38, 35	34, 31, 32	32, 31, 28
DEU	30, 30, 31	38, 40, 33	63, 63. 65	48, 46, 44	63, 65, 58
FRA	34, 33, 31	57, 61, 44	50, 52, 52	75, 74, 77	61, 60, 60

Table 4. Top (+3 SD) bull bias (%) for three GMACE predictions (GM_yng, GM_ms, rGM_ms(v)).

Country	Protein	Stature	SCS	Longevity	Fertility
CAN	21, 22, 10	11, 19, -9	8, 7, 5	62, 61, 44	35, 23, 18
USA	-9, -12, -10	n/a	-0, -2, -3	-11, -17, -10	n/a
NLD	5, 5, 5	2, 4, 3	13, 14, 7	18, 9, 7	37, 28, 18
DEU	10, 9, 10	20, 22, 14	12, 11, 20	5, 1, -5	30, 28, 13
FRA	17, 14, 4	39, 46, 16	17, 18, 19	35, 36, 38	28, 25, 27