Accounting for Residual Correlations among Regional Genomic Predictions via GMACE

P.G. Sullivan
Canadian Diary Network, Guelph, ON, Canada

Abstract

Despite encouraging preliminary results with GMACE, further research has identified some important problems. Errors can be expected in the arbitrary input parameters needed to approximate residual correlations, and GMACE results are quite sensitive to these errors. Restricting bulls to a single GEBV in S-GMACE addresses this concern by eliminating the need for within-sire residual correlations. However, fitting only within-sire residual correlations may be inadequate due to the accumulation of information among relatives via the relationship matrix. It is recommended to further restrict S-GMACE, by preventing the accumulation of genomic information among relatives, to expedite an international genomic evaluation service for young genomically-tested animals. Research should also continue to develop and test less-restricted and/or full-scale GMACE applications.

Key words: genomics, international evaluation, MACE, GMACE, S-GMACE

Introduction

Fitting residual correlations in GMACE (GM) can account for repeated use of the same phenotypic information by different countries for national genomic evaluations of a bull. This strategy prevents multiple counting of the information when national genomic predictions are combined in a global system. The application of residual correlations in the GMACE equations of a single bull was extended to a full population evaluation of genotyped and non-genotyped males and females, including methods for de-regression, international evaluation (VanRaden and Sullivan, 2010) and reliability approximation (Sullivan and VanRaden, 2010). However, several concerns were also identified, so ongoing development and testing of the methods are needed. The purposes of this paper were to provide an update on these developments and to propose an evolution of international genomic evaluation services to be provided by Interbull.

Data

Initial testing of the methods presented by Sullivan and VanRaden (2010) were very encouraging. Results for GMACE seemed nearly equivalent to a global genomic evaluation system, the latter approach being theoretically appealing but currently impractical. The results were based on simulated data with:

1. All animals genotyped.
2. Same definition of GEBV (=DGV) in all countries, ignoring residual polygenics,
3. Regional GEBVs based on multiple-trait models (country=trait),
4. Complete sharing of phenotypes and genotypes within a region, and
5. Little variation among countries in the additional EDCs from genomics for a given bull.

The simulated data were modeled after existing populations of Brown Swiss dairy cattle in 9 countries, and as such was suitable for testing GMACE models and software. However, in terms of genotyping and genomic evaluation, the simulated data were much more representative of a possible future scenario than for current situations in various breeds, where:

1. Not all animals are genotyped, with different patterns of selective and multiple genotyping for young bulls versus ancestors.
2. GEBV definitions vary among countries, particularly in terms of polygenic contributions.

3. DGV methodologies are largely single-trait, with use of MACE proofs of foreign bulls as input.
4. Sharing of phenotypes or genotypes within a region may be incomplete for various reasons.
5. Comparing official GEBV and EBV of a given bull, the additional EDCs from genomics can be quite variable due to the use of MACE proofs for GEBV calculations.

It was not clear if the simulation results could be extrapolated to a GMACE service, due to these important differences between the simulated data and the GEBV input data available to Interbull from member countries. Therefore further studies were conducted using both the above simulated data, and also some data that were submitted to Interbull for a Simplified-GMACE pilot study.

Methods

Correlations in Simulated Data

The previous simulation study demonstrated the potential of GMACE when data are ideal and information is complete. In the present study, errors were introduced to examine the sensitivity of GMACE to sub-optimal input data and parameters.

The relative increase in a bull's EDC due to genomic enhancement (of the bull's EBV going to GEBV) is a critical parameter (γ) in the expected residual correlation among multiple GEBV for the bull. This parameter can be approximated based on genomic relative to traditional reliabilities, but currently there is no standard method to approximate genomic reliability or to ensure it is comparable to the separately approximated traditional reliability. Thus, arbitrary errors are expected in γ and these errors might be different for each country. In the previous study, the assumed values for γ were arbitrary but based on experience of the authors with North American genomic evaluation data (will be referred to as the "correct γ" for discussion purposes). For regional GEBVs γ was higher than for national GEBVs due to the increased information for genomic predictions from regional sharing of data. To test the sensitivity of GMACE to this parameter, the increase in γ due to regional data sharing was ignored (γ-X).

It was known from previous (unpublished) work that large variation in the values of γ for a given bull could be problematic in GMACE. An alternative residual covariance matrix was therefore proposed in GMACE* (GM*) that appeared to reduce the problem. The matrix in GM* included higher residual covariance if there was variation in γ among countries for a bull (Sullivan and VanRaden, 2010). Both GM and GM* were applied in the present study to compare and demonstrate their relative sensitivities to errors in γ.

The proportion of data shared among countries within the same region (parameter *c*) ranges from 0 to 1. Correct values of *c* for the simulation were 1.0 between countries in the same region and 0.0 between countries in different regions. To test the sensitivity of GMACE, parameter *c* was set to either 0.0, 0.5 or 1.0 within a region (always 0.0 between regions).

Data edits for the Interbull Pilot Study

The GMACE software was provided to Interbull for testing on Holstein data, and was verified to produce equivalent results as the Mix99 software (Jakobsen and Jansen, personal communications), for a traditional MACE model and input data. The tests of GMACE by Interbull have so far been limited to a simplified data set (S-GMACE), where GEBVs are restricted to one per bull. The idea behind this restriction was to avoid residual correlations in the data, such that the regular MACE assumption of zero residual correlations among countries would apply. This approach allows for the use of either the GMACE software or Mix99, because it reduces GMACE to a regular MACE application. Details about the Interbull pilot studies on S-GMACE were reported by Zumbach *et al.* (2011). Included in the present paper are additional discussions about the validity of assuming zero residual correlations with this type of S-GMACE approach.

Results and Discussion

Residual Correlation Assumptions

The sensitivities of GMACE (GM) and GM* to parameters affecting residual correlations are shown in Table 1. With the correct γ and *c* parameter values, GM results from the simulated data were nearly BLUP (maximum R^2 and ***b*** close to 1.0). However, with incorrect γ-X, results for GM were much worse than either applying genomics without MACE, or applying MACE without genomics, both in terms of R^2 and ***b***.

Results for GM* with correct γ were not quite as good as GM but still close to BLUP (***b***=0.97 for all sires and 0.93 for young bulls without daughters). With incorrect γ-X, GM* was dramatically better than GM, but still further from BLUP than either genomics without MACE, or MACE without genomics. Considering all tested values of parameter *c*, GM* was more consistent than GM with respect to errors in *c*. In particular, when using a poor value γ-X, results were only consistent for GM*. For example, values of ***b*** for young bulls were respectively 0.61, 0.67 and 0.74 for GM* compared with 0.61, 1.11 and -0.08 for GM with improving values for *c* (0.0, 0.5 and 1.0).

Assumptions that must be made about residual correlations (GM versus GM*), data sharing among countries (parameter *c*) and genomic EDCs (parameter γ) have very important implications for both the reliabilities (R^2) and the scaling (evidenced by ***b***) of international GEBV that would be produced by Interbull, especially for young genotyped bulls. The significant deviations of ***b*** from the expected BLUP value of 1.00 indicate that substantial mis-rankings could occur, especially when comparing young genomically-proven bulls either to older progeny-proven sires with daughters in a single country, or to international sires with daughters in multiple countries.

Data Edits to Avoid Residual Correlations

Restricting sires to a single GEBV as input to S-GMACE can avoid residual correlations for the sire, but it does not prevent accumulation of residual correlations through pedigrees. Interbull review of the pilot results has confirmed this as a valid concern. For sires without daughters but with genotyped sons in multiple countries, reliabilities from S-GMACE were much higher than expected, approaching 100% for a few of these sires. Concerns about upward-bias in reliability for a sire also extend to variance of the sire's EBV, since EBV variance equals genetic variance times reliability. Therefore, it may still be necessary to model residual correlations at the population level even after restricting input data to a single GEBV per bull.

The simulation results in Table 1 (***b***<1.00) are also consistent with an over-scaling of GMACE EBVs, especially for young bulls relative to proven sires. This over-scaling could be explained by a double-counting of genomic information due to inadequate modeling of residual correlations at the population level.

It may be possible to avoid, rather than model residual correlations at the population level, by preventing propagation of genomic information through pedigree relationships within an S-GMACE system. This concept was first suggested as a way to correct approximate reliabilities from S-GMACE, but it can also be applied to the EBVs. International genomic evaluations should be better than regular MACE, which would be the expectation for this new type of approach. The same cannot always be said for GMACE (Table 1) and S-GMACE approaches tested so far. These options to either model or avoid residual correlations at the population level are under investigation.

Variation Explained by Genomics

Genomic predictions currently explain less than 100% of polygenic variance. The proportion explained is a function of methodology, population structure, and the combination of genotypic and phenotypic data available for genomic prediction. All of these factors vary among countries. The relative proportions explained and the unexplained portions that are common among countries affect residual correlations among national GEBVs. This consideration could increase

residual correlations within a region, but would more importantly add non-zero residual correlations between regions into the GMACE system. These concerns should lessen over time as higher density genotyping and imputation methods become more common and effective.

GEBV definitions of each country

Just as the international standardization of trait definitions has affected genetic correlations among countries, the lack of similarity of GEBV definitions and methodologies for DGV and GEBV prediction can affect the residual correlations among countries. It is not yet clear if GEBV definitions can be handled adequately with an approximate methodology (e.g. via residual correlations) or if GEBV covariance estimation will be required for GMACE. Future research may be needed in this area.

Regional versus National De-regression

So far, only national de-regressions have been used within (S-)GMACE. Regional de-regression might be preferred when GEBVs submitted to Interbull are from a regional genomic evaluation system, whereas national de-regression might be preferred when the GEBVs are from national systems, even if regional data were shared among the national evaluation centres. It should be noted that use of regional de-regression would in no way reduce the need to determine and fit an appropriate residual covariance structure among the national populations.

Methods and software for regional deregression are already available as part of the MT-MACE package. A logical next step will be to extend the package to include a multi-trait, genomic option (i.e. MT-GMACE). This would allow for additional modeling options in future research. For example, a multiple-trait application might help to address variation among GEBV definitions in different countries.

Simplifying S-GMACE

The international evaluation service will improve significantly when S-GMACE results are available for young genotyped bulls without daughters. However, sires with progeny proofs already receive quite good international evaluations from regular MACE (Table 1). For these sires, regular MACE results may also be better than GMACE, considering the potential problems with arbitrary parameters used to approximate the residual correlations among national GEBV. To eliminate present concerns about double-counting genomic information, the system can be simplified to focus mainly on using GEBVs for young bulls in combination with regular MACE EBVs for progeny-proven sires. The strategy is to include additional information from genomics (national GEBV – MACE EBV) at the individual level only. For example, this additional genomic information for an individual would be converted among all country-scales but it would not contribute to the MACE EBV of the bull's sire, full-sibs, etc.

Moving towards GMACE

An S-GMACE approach can expedite the availability of an international evaluation service for young genomically-tested bulls. However, there is still significant interest in a full-scale or at least less restrictive service. Therefore, research should continue on the use of residual correlations in GMACE, and suitable methods to eliminate double-counting of genomic information at the population level.

Acknowledgements

Thanks to Paul VanRaden, Gerrit Kistemaker, Jette Jakobsen and Birgit Zumbach for useful comments and discussions.

References

Zumbach, B., Jakobsen, J., Forabosco, F., Jorjani, H. & Dürr, J. 2011. Data selection and pilot run on Simplified Genomic MACE (S-GMACE). Interbull Workshop, Feb 27-28. Guelph, Canada. *Interbull Bulletin 43,* 7-14.

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.

Table 1. Squared correlation (R^2) and regression (b) of true (simulated BV) on predicted international breeding values (EBV).

Sires	Country of most Daughters	[z]Scales of EBV	n	Global Genomics	[y]GMACE options (-X ignores data sharing) GM* c γ	GM c γ	M c-X γ	GM* c γ-X	GM c γ-X	M c-X γ-X	Regional Genomics, No MACE	MACE, No Genomics
Young		All Countries	120	0.60	0.62	0.61	0.61	0.62	0.02	0.59	0.52	0.13
1st Crop			1476	0.68	0.65	0.66	0.63	0.64	0.23	0.63	0.52	0.61
All			8193	0.67	0.64	0.65	0.62	0.64	0.28	0.62	0.48	0.62
Daughters in 1 country	Region 1	Local Region 1	2108	0.74	0.72	0.72	0.71	0.71	0.35	0.69	0.72	0.69
	Region 2	Local Region 2	5003	0.70	0.69	0.69	0.68	0.68	0.33	0.66	0.70	0.66
	Region 1	Foreign Region 2	2108	0.71	0.67	0.68	0.62	0.68	0.32	0.64	0.37	0.66
	Region 2	Foreign Region 1	5003	0.66	0.60	0.62	0.56	0.61	0.21	0.59	0.33	0.61
Daughters in multiple countries	Region 1	Local Region 1	386	0.84	0.81	0.82	0.80	0.81	0.59	0.79	0.80	0.82
	Region 2	Local Region 2	468	0.84	0.82	0.82	0.81	0.81	0.68	0.79	0.81	0.81
	Region 1	Foreign Region 2	386	0.81	0.79	0.80	0.78	0.79	0.61	0.77	0.70	0.79
	Region 2	Foreign Region 1	468	0.80	0.76	0.78	0.73	0.77	0.54	0.75	0.57	0.78
Young		All Countries	120	0.95	0.93	1.00	0.86	0.74	-0.08	0.61	0.97	0.70
1st Crop			1476	0.99	0.95	1.00	0.90	0.81	0.79	0.71	1.01	0.94
All			8193	0.99	0.97	1.01	0.92	0.83	0.87	0.73	0.94	0.95
Daughters in 1 country	Region 1	Local Region 1	2108	1.00	0.96	1.01	0.91	0.83	0.85	0.75	0.98	0.95
	Region 2	Local Region 2	5003	1.02	1.00	1.04	0.94	0.86	0.94	0.75	1.01	0.98
	Region 1	Foreign Region 2	2108	1.00	1.00	1.04	0.94	0.86	1.08	0.76	0.91	0.94
	Region 2	Foreign Region 1	5003	1.01	1.00	1.05	0.94	0.83	0.75	0.73	0.91	0.96
Daughters in multiple countries	Region 1	Local Region 1	386	1.01	0.95	0.99	0.90	0.86	1.00	0.79	0.99	0.99
	Region 2	Local Region 2	468	1.02	0.98	1.02	0.94	0.90	1.13	0.83	1.01	1.01
	Region 1	Foreign Region 2	386	1.00	0.95	0.99	0.90	0.87	1.09	0.79	1.02	0.99
	Region 2	Foreign Region 1	468	1.00	0.98	1.02	0.93	0.87	1.01	0.80	0.99	0.98

[z]Results presented were simple averages of the country results within a region. Region 1 results excluded New Zealand due to limited data.
[y]M(ACE), GM(ACE) and GM(ACE)* defined in Sullivan and Van Raden, 2010