Optimized Aggregation of Phenotypes for MA-BLUP Evaluation in German Fleckvieh

Edel, C., Emmerling, R. and Götz, K.-U.

Bavarian State Research Center for Agriculture, Institute of Animal Breeding, D-85586 Grub/Poing, Germany

Abstract

This paper presents some of the methodological aspects in the aggregation of phenotypes for a classical application of marker assisted breeding value estimation in German Fleckvieh ('InfraMAS'). The methods used are partly illustrated and include the calculation of YD based on regression functions and 'best prediction', the calculation of DYD for bulls and bulldams and the calculation of the weighting factors and matrices, respectively, that are necessary in combining these different kinds of information. A possible strategy for situations where in a multiple-trait setting one trait is systematically missing and the respective YD/DYD is not defined is proposed and shortly described. Results from validation studies during the developmental process are shown. They suggest that the information content of the complete routine evaluation can be reconstructed to a large extent by using the proposed aggregation strategy.

Keywords: phenotype aggregation, marker-assisted selection, DYD, weighting factor

1. Introduction

During the last two years the Institute of Animal Breeding of the Bavarian State Research Center in Grub (ITZ) has developed a marker-assisted genetic evaluation system (MA-BLUP) for Fleckvieh that is now routinely conducted since the beginning of 2009 ('InfraMAS'). From a today's perspective the classical approach of marker assisted breeding value estimation seems to be by far less flexible and general than what is promised by 'genomic selection'. The late introduction of a seemingly obsolete method might from this point be questioned. However, besides possible benefits in the selection process, developing and introducing this application had beneficial side-effects for us as well as for the breeders and AI-stations involved. Research with respect to methodological aspects, clarification of open questions of logistics and workflow, the distribution of costs and redistribution of results and the practical use of this additional information in the existing selection process may be of great value in the development of applications of genomic selection for German Fleckvieh that are already underway.

In this paper we present how we use and aggregate phenotypic information in our MA-BLUP application. The investigations are focused on strategies using daughter yield deviations (DYD, VanRaden and Wiggans, 1991) for bulls and DYD and yield deviations (YD) of cows. The application developed for the estimation of variance components and the MA-BLUP evaluations is based on a traditional two- step approach (George *et al.*, 2000; Szyda *et al.*, 2005, Druet *et al.*, 2006).

Neuner *et al.* (2008) showed that in two-step approaches there is a severe loss in accuracy due to unaccounted phenotypic information, especially of untyped bulldams. In some cases this might lead to a situation, where the use of QTL-information and MA-BLUP is not beneficial at all. In German Fleckvieh a large percentage of selected bull dams are in fact genotyped and therefore the

inclusion of their phenotypes is desirable. Since the amount of additive-genetic variance explained by markers is limited, it is important to restore the information content of the full routine-evaluation as completely as possible.

2. Methods

Starting from the standard decomposition of the estimated breeding value, we want to illustrate the succeeding steps of aggregation. The standard method was developed for our multivariate random-regression test-day models for milk production traits (Emmerling *et al.*, 2002). Some additional considerations are included dealing with the problem of standard multivariate models with many missing values.

A standard calculation scheme proceeds as follows:

- Calulation of YD and weights of YD for cows.
- Calculation of DYD and the weights of DYD for bulls and bulldams.
- Phenotypes and weights are used in the MA-BLUP evaluation.

YD of daughters which are themselves part of the system (genotyped) are left out in the calculation of DYD for bulls and bulldams.

2.1 Steps of aggregation

To illustrate the most important steps involved, we will start from a well known approximate decomposition of a vector of estimated multivariate breeding values (Mrode and Swanson, 2004):

$$\hat{\mathbf{a}}_{anim} = \mathbf{M}_1(\mathbf{PA}) + \mathbf{M}_2(\mathbf{YD}) + \mathbf{M}_3(\mathbf{DYD}),$$

where $\mathbf{PA}$ is a vector of parent averages, $\mathbf{YD}$ is a vector of performances of the animal corrected for all effects other than additive-genetic, $\mathbf{DYD}$ is a vector of daughter yield deviations and $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ are weighting matrices.

2.1.1 Calculation of YD

The mixed model equations solver MiX99 (Lidauer *et al.,* 1999), routinely provides testday-observations of cows corrected for model terms that can be specified by the user. Lactation yield deviations are then computed in two steps, first by calculating YD expressed as (regression-)functions ($\mathbf{YD}_F$) by standard formula

$$\mathbf{YD}_F = (\mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q})^{-1}(\mathbf{Q}'\mathbf{R}^{-1}\tilde{\mathbf{y}}),$$

where $\tilde{\mathbf{y}}$ is a vector of pre-corrected testday performances and $\mathbf{Q}$ is a matrix containing the covariables corresponding to the observed DIM. In a second step lactation yield deviations are calculated by taking columnwise sums over testday covariables for each lactation in $\mathbf{\Phi}$, here defined as a matrix containing the covariables of all possible testdays. For example the first lactation YD (in a setting only considering the first lactation) then would be

$$\mathbf{YD}_{1L} = \mathbf{\Phi}_{[6:310,:]}\mathbf{YD}_F .$$

Alternatively, we calculate lactation yield deviations based on best prediction (VanRaden, 1997) using

$$\mathbf{YD}_{BP} = \mathbf{1}'\mathbf{C}\mathbf{V}^{-1}\tilde{\mathbf{y}},$$

where $\mathbf{V}$ is the phenotypic (co)variance matrix of the observed testdays, $\mathbf{C}$ is the (co)variance matrix between observed testdays and all possible testdays and $\mathbf{1}$ is a vector containing 1 for the days of the specific lactation and zero otherwise.

For the calculation of DYD the YD of daughters are corrected for the breeding value of the mate. This can either be done on the

level of YD-functions or on the level of lactation YD. All following calculations (YD, DYD of cows) are done on the level of lactation YD.

2.1.2 Calculation of DYD

DYD are calculated by

$$\mathbf{DYD} = (\sum u_{ik}\mathbf{N}_k)^{-1}\sum u_{ik}\mathbf{N}_k(2\mathbf{YD}_k - \hat{\mathbf{a}}_{mk})$$

and are basically a weighted mean of k daughter $\mathbf{YD}$ corrected for the breeding value of the respective dam of the daughter ($\hat{\mathbf{a}}_{mk}$) where each one of the k $\mathbf{YD}$ is weighted by a term that is a combination of the reliability of the daughter's performance ($\mathbf{N}_k$) and a term that accounts for the covariance of mendelian sampling terms (u_{ik}). Leaving out the k subscript for better readability

$$\mathbf{N} = (\mathbf{Z}\mathbf{R}^{-1}\mathbf{Z} + u_{ii}\mathbf{G}^{-1})^{-1}(\mathbf{Z}\mathbf{R}^{-1}\mathbf{Z}),$$

where $\mathbf{Z}$, $\mathbf{R}$ and $\mathbf{G}$ are taken from standard-nomenclature. It can be shown that for $u_{ii} = 1$

$$\mathbf{N} = \Re',$$

where $\Re$ is a standard reliability matrix (Liu *et al.*, 2004). It should be mentioned that matrix $\mathbf{N}$ (or $\Re$) is non-symmetric.

Processing each bull or cow in turn, the $\mathbf{YD}$ of all daughters that are not themselves part of the MA-BLUP system are collected and corrected for the breeding value of the specific mate. The $\mathbf{N}$ matrices and u_{ik} terms are calculated and stored and the final weighted mean is calculated. All these calculations are done on lactation level and are therefore approximate.

2.1.3 Calculation of weights

In our application we use the effective number of own performances (EOP) as weights of DYD and YD, a measure that is closely related to the *effective daughter contribution* (EDC) concept (Fikse & Banos, 2001; Liu *et al.*, 2001). In a multivariate setting an EOP would be

$$\varphi_{EOP} = \left[(\mathbf{I} - \Re)^{-1} - \mathbf{I}\right]\mathbf{G}^{-1}\mathbf{R},$$

where $\Re$ again is the (approximate) standard reliability matrix (Liu *et al.*, 2004). Analogy to the EDC concept of Liu *et al.* (2001) can be illustrated by showing that an EDC would be

$$\varphi_{EDC} = \left[(\mathbf{I} - \Re)^{-1} - \mathbf{I}\right]\mathbf{G_S}^{-1}\mathbf{R_S},$$

where $\mathbf{G_S} = 0.25\mathbf{G}$, $\mathbf{R_S} = \mathbf{P} - 0.25\mathbf{G}$ and $\mathbf{P}$ is the phenotypic (co)variance matrix. Again φ_{EDC} and φ_{EOP} in this formulation are matrices that are not symmetric. For storage purposes a symmetric from can be found by leaving out the multiplication by $\mathbf{R}$ (or $\mathbf{R_S}$).

2.1.4 Aggregation to a single trait

To handle nonsymmetrical matrices as weights in parameter or breeding value estimation we tested different forms of linearization (e.g. Sullivan, 2002). Finally we decided to build a single synthetic trait in analogy to the selection criterion, which is in our case an unweighted mean of the breeding values of the first, second and all higher lactations. For the aggregation we use a simple selection-index approach applied after the calculation of multi-trait DYD and YD, respectively, where

$$DYD_{tot} = \mathbf{b}'\mathbf{DYD} * \frac{1}{h^2_{tot}}$$

with $\mathbf{b} = \mathbf{P}^{-1}\mathbf{Ga}$, $\mathbf{a} = [1/3, 1/3, 1/3]$ and

$$h^2_{tot} = \frac{\mathbf{a'Ga}}{\mathbf{a'Pa}}$$

Accordingly, scalar weights had to be calculated, which was done by using a strategy outlined in the Interbull code of practice (2004) resulting in a scalar reliability for each information source (ω). This reliability is finally converted to EOP by

$$\varphi_{EOP} = \frac{\lambda}{1-\omega} - \lambda ,$$

where $\lambda = \mathbf{a'Ra}/\mathbf{a'Ga}$.

2.2 Additional aspects of application

There might be a situation in a multiple-trait-scenario where a large proportion of animals systematically have observations on one trait only. Consider two countries with different systems of measurement of the same trait and a common genetic evaluation system that treats both traits in a multivariate setting, as it is the case for the trait milkability in Germany and Austria. The selection criterion might then be one of the two traits or a linear combination of both. Using DYD and YD in such a situation has the drawback that if all daughters of a bull have only observations in one of both traits, the DYD of the other trait is not defined and this form of correlated information is lost. In cases where this loss is not acceptable and the correlated information has to be recovered we follow all steps outlined so far, but for each information source breeding values are estimated using pre-corrected observations and a selection index approach. The estimated breeding values are then aggregated and are finally de-regressed again by using a scalar reliability connected with the respective amount and sort of information.

In our MA-BLUP system we do not include parent averages from the routine evaluation. The pedigree consisting entirely of genotyped animals is routinely extended by one generation of ungenotyped animals whose genotypes are inferred. LOKI (Heath, 1997), the software that is used for the calculation of the QTL-IBD matrix is able to infer genotypes of ancestors to some extent. Including the DYD and YD of these ungenotyped ancestors compensates largely for not including the PA component in the calculation of the breeding values of genotyped animals.

2.3 Performance and Verification

To verify the performance of the applied weighting and aggregation strategies we calculated correlations between the breeding values estimated based on aggregated phenotypes to those from the standard routine evaluation (all information used was from the joint German-Austrian routine evaluation, April 2008).

3. Results and Discussion

Table 1 summarizes the results for the trait milk yield. Increasing the available information, starting from DYD of bulls (DYDb) only, to an additional inclusion of the YD of bulldams and finally the DYD of bulldams (DYDc), results in a growing correlation to the breeding value from the routine evaluation, especially for cows and candidates. This confirms the results of Neuner *et al.* (2008) and emphasizes, that in a two-step approach all sources of information for the estimation of the breeding value of an animal should be included.

The last column illustrates that inclusion of YD based on ‘best prediction’ (YDbp, also used for the calculation of DYDc) instead of

using YD based on individually estimated regression functions leads to a substantially higher correlation for bulldams. Further analyses showed that estimating individual regression functions solely from the few observations of an animal, without the stabilizing effect of the pedigree, leads in many cases to unreasonable functions, heavily influenced by a few measurements. Using best prediction in this context is more robust with respect to outliers and influential observations.

Correlations of .99, .96 and .97 for old bulls, cows and candidates, respectively, give some evidence that a large percentage of the information of the complete routine evaluation can in fact be restored by using relatively simple aggregating techniques. Estimates of QTL-effects derived from such a system should be largely unbiased and as precise as possible. However, in routine evaluation a correlation of .97 leads to a large amount of re-ranking between candidates caused only by loss of information. Therefore, in our current routine application we conduct two estimations, one without using QTL information and one including it to derive the additional effect of QTL given the information of the subsystem. This 'QTL-contribution' is finally combined with the breeding value from standard routine evaluation to give the final ranking criterion. This is similar to the way genomic information is considered in the Dutch breeding program for Holsteins (De Roos *et al.*, 2009).

4. Acknowledgement

This work and the development of InfraMAS were financed by the Bavarian Ministry of Agriculture and supported by breeding organizations and AI stations of German Fleckvieh in Bavaria.

Table 1. Results from a validation study. Figures shown are correlations between breeding values of the routine evaluation for milk yield (April 2008) and breeding values estimated from a subsystem of genotyped animals and aggregated phenotypes. DYDb is a subsystem including the DYD of bulls only. DYDb/YD additionally includes the YD of bulldams and DYDb/YD/DYDc additionally DYD of bulldams. Finally in DYDb/YDbp/DYDc the YD used are calculated by 'best prediction'.

	Nobs	**DYDb**	**DYDb/YD**	**DYDb/YD/DYDc**	**DYDb/YDbp/DYDc**
all bulls	1821	0.969	0.974	0.974	0.974
old bulls	932	0.989	0.990	0.990	0.990
cows	1183	0.585	0.856	0.899	0.959
candidates	258	0.879	0.950	0.954	0.973

5. References

Druet, T., Fritz, S., Boichard, D. & Colleau, J.J. 2006. Estimation of Genetic Parameters for Quantitative Trait Loci for Dairy Traits in the French Holstein Population. *J. Dairy Sci. 89*, 4070-4076.

De Roos, A.P.W, Schrooten, C. Mullaart, E., van der Beek, S., de Jong, G. & Voskamp, W. 2009. Genomic Selection at CRV. *Interbull Bulletin 39,* 47-50.

Emmerling, R., Lidauer, M. & Mäntysaari, E.A. 2002. Multiple lactation random regression test-day model for Simmental and Brown Swiss in Germany and Austria. Proc. Interbull Meeting, Interlaaken, Switzerland. Int. Bull Evaluation Service, *Interbull Bulletin 29*, 111-117.

Fikse, W.F. & Banos, G. 2001. Weighting factors of Sire Daughter Information in International Genetic Evaluations. *J. Dairy Sci. 84*, 1759-1767.

George, A.W., Visscher, P.M. & Haley, C.S. 2000: Mapping Quantitative Trait Loci in Complex Pedigrees: A Two-Step Variance Component Approach. *Genetics 156,* 2081-2092.

Heath, S.C. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet. 61,* 748-760.

Interbull Code of Practice (updated April 27, 2004). Weighting factors for the international genetic evaluation. *Appendix IV,* 1-5.

Lidauer, M.I., Strandén, I. & Mäntysaari, E. 1999. *Mixed Model Equations Solver. MiX99 Manual.* Animal Production Research, Jokioinen, Finland.

Liu, Z., Reinhardt, F. & Reents, R. 2001. The Effective Daughter Contribution Concept Applied to Multiple Trait Models For Approximating Reliabilityof Estimated Breeding Values. *Interbull Bulletin 27,* 41-47.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. Derivation and Calculation of Approximate Reliabilities and Daughter Yield-Deviations of a Random Regression Test-Day Model for Genetic Evaluation of Dairy Cattle. *J. Dairy Sci. 87,* 1896-1907.

Mrode, R.A. & Swanson, G.J.T. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. *Livest. Prod. Sci. 86,* 253-260.

Neuner, S., Emmerling, R., Thaller, G. & Goetz, K.-U. 2008. Strategies for Estimating QTL Variance Components in Marker-Assisted Best Linear Unbiased Predictor Models in Dairy Cattle. *J. Dairy Sci. 91,* 4344-4354.

Sullivan, P. 2002. Genetic evaluation strategies for multiple traits and countries. *Doctoral thesis*. University of Guelph, May, 2002.

Szyda, J., Liu, Z., Reinhardt, F. & Reents, R. 2005. Estimation of Quantitative Trait Loci Parameters for Milk Production Traits in German Holstein Dairy Cattle Population. *J. Dairy Sci. 88,* 356-367.

VanRaden, P.M. & Wiggans, G.R. 1991. Derivation, calculation and use of national animal information. *J. Dairy Sci. 74,* 2737-2746.

VanRaden, P.M. 1997. Lactation yields and accuracies computed from test day yields and (co)variances by best prediction. *J. Dairy Sci. 80,* 3015-3022.