Revision of random regression test-day model has improved genomic prediction for Nordic Red dairy cattle

T. Pitkänen[1], M.H. Lidauer[1], J. Pösö[2], W.F. Fikse[3], U.S. Nielsen[4], G. P. Aamand[5] and M. Koivula[1]

1Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland

2 Faba co-op, P.O. BOX 40, FIN-01301 Vantaa, Finland

3Växa, Box 288, 75105 Uppsala, Sweden

4 SEGES Innovation, Udkærsvej 15, 8200 Aarhus N, Denmark

5 NAV Nordic Cattle Genetic Evaluation, 8200 Aarhus N, Denmark

Corresponding author: timo.j.pitkanen@luke.fi

Abstract

In 2006, Denmark, Finland, and Sweden have introduced across-country genetic evaluations for yield traits for their Holstein, Red dairy cattle (RCD) and Jersey populations. The implemented breed-specific random regression test-day models (RRM) were the outcome of an intensive research cooperation among the countries. Especially developing the RRM for the RDC population presented unique challenges due to its heterogeneous population structure spanning across Finland, Sweden, and Denmark. As genomic prediction became a key tool for breeding decisions, it became evident that the reliability of genomic enhanced breeding values (GEBV) for the RDC breed was lower than expected when compared to the Holstein and Jersey breeds. Several factors contributed to this discrepancy. Notably, during the last two decades, changes in herd and population structures were most pronounced within the RDC breed where the original RDC country-subpopulations have become much more alike. Thus, revision of the RRM is crucial in enhancing the reliability of GEBV for the selection of breeding candidate animals. First important improvements were the revision of modelling automated milking system data and a newly estimated set of variance components with lower h^2. The updates made so far indicate considerable improvement in the genomic predictions. The LR regression coefficient (b_1) values increased for example for milk yield from 0.85 of the original model to 0.92 of the revised model, indicating that the bias decreased with the revised model. Also, the coefficient of correlation (R^2) increased for all production traits on average 4.5%. In a next step, we will truncate the phenotypic data, optimize the pedigree information, and study whether modelling metafounders for the heterogeneous RDC population will result in further improvements for the genomic prediction.

Key words: Nordic Red dairy cattle, yield evaluation, automatic milking system

Introduction

In 2006, the Nordic countries, Finland (FIN), Denmark (DNK), and Sweden (SWE), introduced across-country genetic yield evaluations for their dairy cattle populations. The outcome of intensive research cooperation was the random regression test-day models (RRM) for Holstein, Red dairy cattle (RDC), and Jersey. In 2010 the models were updated to replace the Swedish lactation yield observations by test-day observations and to apply a common set of variance components instead of country-specific variance components (Lidauer et al., 2015) and continue to serve as the basis for predicting genomic enhanced breeding values (GEBV).

From the beginning, the model for the RDC breed proved to be the most challenging due to the heterogeneous population structure of RDC cattle across the Nordic countries. After genomic predictions for the three breeds were built, it was observed that the reliability of GEBVs for RDC was lower than expected, especially when compared to the reliability of GEBVs for Holstein and Jersey. Potential reasons for this discrepancy include changes in herd and population structures, which may have made the applied RRM suboptimal for genomic prediction. Revising the RRM for RDC

evaluation was found to be crucial for improving the reliability of GEBVs and for the optimal selection of candidate animals.

Materials and Methods

Model

The applied multiple-trait RRM describes test-day milk, protein, and fat yields for a cow's first three lactations using nine model equations. Each trait has a random regression function for random genetic and permanent environmental effects. The model includes fixed effects nested within countries, some of them also nested within breed, heterosis and recombination loss adjustments, and adjustments for heterogeneous variance (HV). Due to modelling of covariance functions, 15 cow-specific coefficients define all nine breeding value curves. For more information, see Lidauer et al. (2015).

Revising the variance components

In the old RRM (in use until November 2023), heritability was based on variance components estimated from Swedish data and were considered too high, especially for later parities. Most of the RDC test-day data come from Finland, where earlier studies found lower heritability (h^2) values compared to those used in the RRM. Specifically, variance components for permanent environment and genetic effects in later lactations were too high. This became
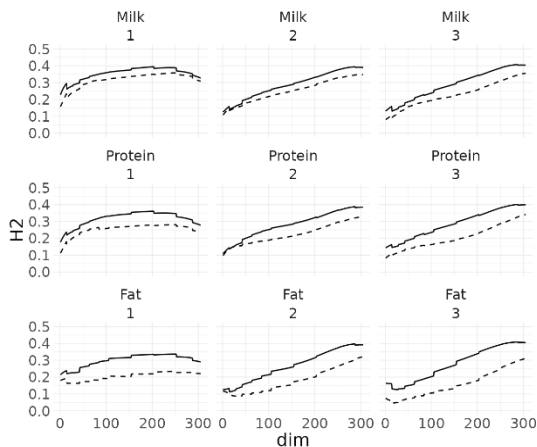

Figure 1. Heritabilities in old (solid line) versus updated (dashed line) model for milk, protein, and fat in lactations 1-3
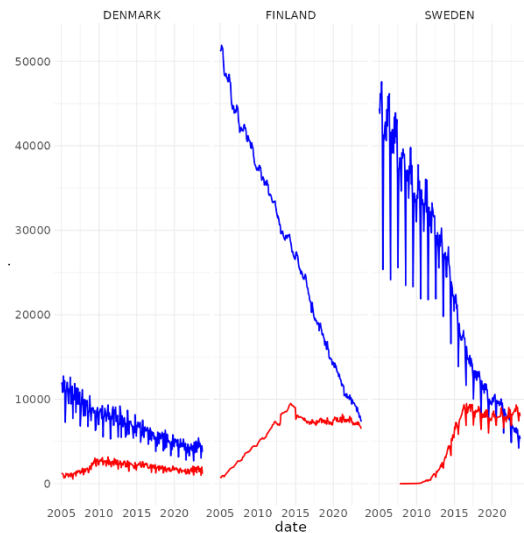

Figure 2. Number of test-day observations for first lactation milk yields by country and milking system. Blue lines from conventional milking system (CMS) and red lines from automatic milking system (AMS).

even more critical after

assigning higher weights to later lactations in the Nordic Total Merit Index. The updated variance components are now based on Finnish data, resulting in lower heritabilities, particularly for protein and fat yields in later lactations (Figure 1). Consequently, now the variance components better fit the data.

Updated modelling of automatic milking system data

Over the last two decades, there has been a rapid decrease in conventional milking system (CMS) observations. Meanwhile, automatic milking system (AMS) observations have increased slightly, and currently, approximately half of the test-day records in Finland, Sweden, and Denmark come from AMS (Figure 2).

The old RRM assumed the same residual variance for all AMS observations. However, in Finland, the measurement protocol has changed twice over the years, with the latest methods based on a smaller number of AMS milkings. The residual variances differ between measurement. This caused issues in simultaneously adjusting for HV, as it overcorrected observations made using the new measurement protocols due to incorrect
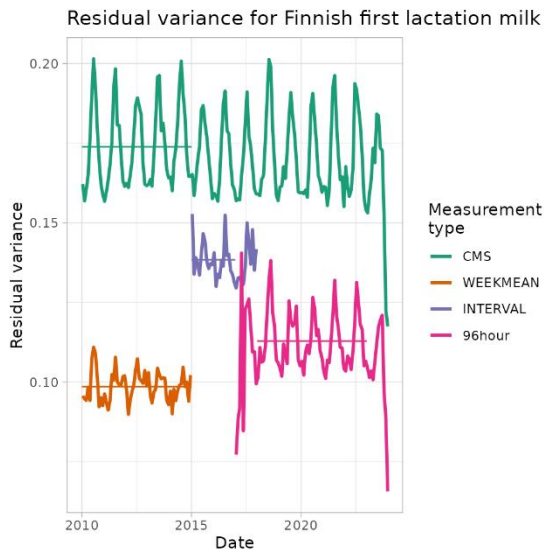
Figure 3. The level of residual variance in different milking system protocols by months from 2010 to 2023. Green line conventional milking system (CMS).

baseline variances. To address this, the model was updated to provide separate residual variances for the different methods to calculated 24h yield from AMS.

The updated milking protocols in Finland are defined 1) as sum of morning and evening milking yields in CMS for whole data collection period. For AMS there are three measurement types: 2) the average of one weeks milkings during years 2010-2015 (weekmean), 3) sum of two successive milkings scaled to 24h yield during 2015-2017 (Interval), and 4) four days average milkings (96hour) from 2017 onwards.

The Figure 3 shows the level of residual variance in different daily milk yield measurements. It shows that the residual variance for CMS is always higher than that for different AMS systems. The correct measurement information for AMS observations was included also in the data, and the HV adjustment was updated to handle the different AMS recording protocols.

Holstein observations in the Finnish test-day data

In the past, Finnish herd sizes were small. As a result, test-day records for Holstein cows were included in the model to increase contemporary group sizes, although the results of the Holstein evaluations were not used from this model.

However, herd sizes in Finland have increased and this is no longer as relevant, so Holstein observations were removed from the Finnish test-day data. This removal also required revising model effects that included breed interactions.

Testing with updated model

All test-day data and genotype data available in February 2023 were obtained from the NAV. The data included 4.7 million cows with observations and 6.2 million animals in the pedigree. Genotype data included 229,706 genotyped animals.

The animal model and single-step RRM were solved by preconditioned conjugate gradient method (Strandén and Lidauer, 1999). The genomic evaluation was realized by ssGTaBLUP (Mäntysaari et al., 2017). For setting up the genomic relationship matrix ($\mathbf{G}$) the VanRaden method 1 and a 10% residual polygenic proportion (RPG) were used, and diagonal of $\mathbf{G}$ was scaled to be on average equal to the pedigree-based relationship matrix of the genotyped animals ($\mathbf{A}_{22}$) (Vanderplas et al., 2023). The pedigree inbreeding coefficients were accounted for both in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$.

Model comparisons were based on forward prediction validation, utilizing solutions from both full-data and reduced-data evaluations. The reduced data were derived by removing the last four years of observations from the full data set. The linear regression validation (LR) method (Legarra and Reverter, 2018) was employed for validation. This method compares predictions based on reduced and full data, yielding estimates of accuracy and bias.

Danish, Finnish, and Swedish bulls born between 2014 and 2018, each having at least 20 daughters in the full-data set but no daughters in the reduced-data set, were defined as candidate bulls. This criterion resulted in a total of 222 candidate bulls.

Results & Discussion

The updated model changed the estimated breeding values (EBV) and the GEBV of the animals, with the correlation between the old and new EBVs averaging 0.97 and 0.98 for GEBVs. Consequently, the RRM update caused some reranking of the bulls. This was expected, as the incorrect AMS protocol had caused issues, particularly for Finnish cows, which also affected bull evaluations. With the corrected milking system data, the (G)EBVs for some bulls changed accordingly.

The comparison of EBVs and GEBVs from the updated model showed that validation of the bulls improved with corrections (Table 1). In genomic animal model bias decreased considerably and both regression coefficient (b_1) and coefficient of determination (R^2) improved compared to the old RRM.

Table 1. Linear regression (LR) results for the validation bulls' breeding values based on the BLUP (EBV) or ssGTaBLUP (GEBV) evaluations applying the old (old) or the updated (new) model. The values in the table are: b0= mean(Full_((G)EBV–reduced_(G)EBV) ±SD, b1 regression coefficient and R2 coefficient of determination.

	Model	b_0	b_1	R^2
Milk	EBV_{old}	129.46 (±540.0)	0.68	0.22
	EBV_{new}	-81.18 (±569.9)	0.69	0.23
	$GEBV_{old}$	-473.88 (±363.7)	0.85	0.66
	$GEBV_{new}$	-291.34 (±354.9)	0.92	0.70
Protein	EBV_{old}	-0.72 (±16.2)	0.62	0.21
	EBV_{new}	1.24 (±17.0)	0.63	0.21
	$GEBV_{old}$	-17.78 (±12.3)	0.75	0.60
	$GEBV_{new}$	-12.41 (±11.9)	0.84	0.63
Fat	EBV_{old}	-1.77 (±21.4)	0.77	0.28
	EBV_{new}	3.20 (±27.9)	0.76	0.27
	$GEBV_{old}$	-22.91 (±15.6)	0.84	0.63
	$GEBV_{new}$	-18.61 (±19.6)	0.91	0.65

The protein genetic trends for bulls are shown in Figure 4A. After the introduction of genomic selection, the genetic trend in single step evaluation started to be much higher in the
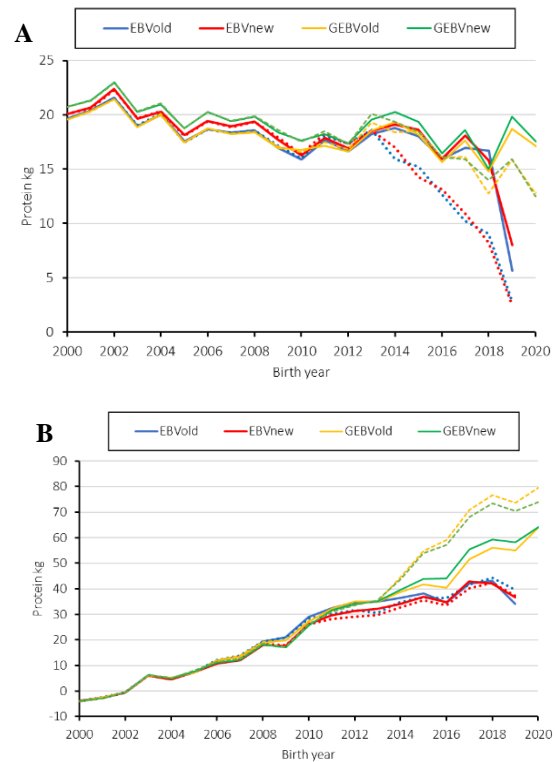

Figure 4. **A**) Genetic trends for protein yield (kg) by birth year averages. (**B**) SD for protein yield breeding values (kg) by birth year. EBVold the old RRM model, GEBVold the single step with old RRM, EBVnew the updated RRM and the GEBVnew the single step with updated RRM. Solid lines are for full data and dashed lines for reduced run.

reduced data compared to that in the full data run with the old RRM, whereas with the updated RRM the difference between the genetic trend in the reduced run and full run is reduced and thus bias is decreased. As expected, the SD of the EBVs and GEBVs slightly increased with the updated model due to new AMS protocol (Figure 4B).

The updates made to the RRM so far have already enhanced the genomic predictions for the RDC. Specifically, the bias has decreased, and the b_1 value has increased. Further changes are underway, including data truncation to exclude test-day records prior to 2005 and optimization of pedigree. Additionally, work is ongoing to refine the definitions of calving age, heterosis, and recombination effects. Once all effects in the RRM have been updated, we will investigate replacing unknown parent groups

with metafounders to assess whether this approach further improves genomic predictions.

Conclusions

As a final remark, changes made so far in the model improved validation results. These changes included updating variance components, improved handling of residual variance of Finnish AMS records, as well as removing Holstein observations.

Acknowledgements

References

Legarra, A., Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol. 50*, 53. doi: 10.1186/s12711-018-0426-6.

Lidauer, M., Pösö, J., Pederson, J., Lassen, J., Madsen, P., Mäntysaari, E.A., Nielsen, U., Eriksson, J-Å., Johansson, K., Pitkänen, T., Strandén, I., Aamand, G.P. 2015. Across-country test-day model evaluations for Nordic Holstein, Red Cattle and Jersey. *J. Dairy Sci. 98*, 1296–1309. doi: 10.3168/jds.2014-8307.

Mäntysaari, E.A., Evans, R.D., Strandén, I. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J. Anim. Sci. 95*, 4728-4737. doi: 10.2527/jas2017.1912.

Strandén, I., Lidauer, M 1999. Solving large mixed models using preconditioned conjugate gradient iteration. *J. Dairy Sci.* 82, 2779-2787. doi: 10.3168/jds.S0022-0302(99)75535-9.

Vandenplas, J., ten Napel, J., Darbaghshahi, S.N., Evans, R., Calus, M.P.L., Veerkamp, R., Cromie, A., Mäntysaari, E.A.; Strandén, I. 2023. Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet. Sel. Evol* 55, 37. Doi:10.1186/s12711-023-00808-z

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91*, 4414-4423. doi: 10.3168/jds.2007-0980.