Differential handling of missing parents in genetic evaluation of dairy cattle using single-step test-day SNP-BLUP model

Dawid Słomian[1], Kacper Żukowski[1], Monika Skarwecka[1], Jan Ten Napel[3], Jeremie Vandenplas[3] Joanna Szyda[1,2]

[1]National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

[2]Biostatistic Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland

[3]Animal Breeding and Genomics, Wageningen University & Research, P. O. Box 338, 6700 AH Wageningen, Netherlands

Abstract

Single-step genomic models use all available information on animals' phenotype, genotype and pedigree. Nowadays, many countries aim towards implementing single-step models and replacing the existing conventional models for routine evaluation. Even in the area of genomic evaluation, the pedigree data has still a significant impact on estimated genomic breeding values, and therefore it is very important to obtain the most informative structure of the pedigree. The crucial aspect of the pedigree editing is handling missing parents information. Missing data can arise either due to truly missing parentage information, or due to the fact that not all generations are utilized. We focused on three scenarios for handling missing parents: 1) raw pedigree, where missing parents IDs were set to missing; 2) genetic groups, where missing parents in the raw pedigree were replaced by genetic groups based on year of birth, country of origin, and sex; 3) metafounders, which are created based on genetic groups and genomic information. The genomic breeding values for fat yield were estimated using the single-step test-day SNP-BLUP model implemented by the MiXBLUP software. The analysed data corresponds to the population of Polish Holstein Friesian cattle used for routine genetic evaluation. We compared the results of the validation obtained by the three pedigree handling approaches and observed that the best results of validation were achieved by the scenario with metafounders (3), followed by scenario fitting pedigree with genetic groups (2), and finally by the raw pedigree (1). The metafounders scenario uses most of the information including genotype data, therefore, it provides the best classification of unknown animals into groups, which improves validation results.

Key words: single-step models, genetic groups, metafounders, validation

Introduction

The structure of pedigree data is important for the routine genetic and genomic evaluations of dairy cattle (Bradford et al., 2019). To reduce the amount of missing data and the corresponding bias in the pedigree file, genetic groups (phantom parents) are used to divide the missing ancestors into different categories (Westell et al., 1988, Legarra et al., 2007). Nowadays, single-step genomic models are the models of choice of many countries that are working on implementing routine breeding value evaluation. The single-step model incorporates all available sources of information, i.e., phenotype, genotype, and pedigree.

In this study, we focused on a single-step random regression SNP-BLUP test-day model for fat yield and investigated three approaches to handle missing parents in a pedigree: 1) *raw pedigree* with missing parents IDs set to missing; 2) *genetic groups* with missing parents replaced by unrelated genetic groups, which are defined based on year of birth, sex, and country of origin; 3) *metafounders* with missing parents replaced by metafounders, which are genetic groups with

relationships estimated from genomic information of descendants. The goal of this study was to compare results of genetic trend validation, number of iterations required to estimate all solutions, and computing times of the single-step evaluations with the three different pedigree handling scenarios. We also compared the results of the conventional pedigree-based BLUP (single-trait random regression test-day BLUP) with or without genetic groups with the single-step random regression test-day SNP-BLUP.

Materials and Methods

The data set (Table 1) corresponds to the Polish national evaluation for fat yield from April 2024 and contains 63,484,231 records of 3,701,610 cows in full data set, and 58,441,242 records of 3,224,577 cows in the truncated data set with the individuals born from 2019 removed. Genomic information from 46,118 SNPs was available for 182,143 animals. Pedigree information included 4,513,226 individuals and was extracted up to the third generation from animals with phenotypes or genotypes.

Table 1: Number of animals in the analysed data set.

Data	Sex	Number of animals	Number of records
Phenotype (fat yield)	Cows	3,701,610	63,484,231 Full data set 58,441,242 Truncated data set
Genotype	Cows Bulls	113,171 68,972	182,143
Pedigree	Cows Bulls	4,418,710 94,516	4,513,226

Genetic groups were defined according to the year of birth, sex, and country of origin of the animals with at least one missing parent (Table 2). All animals born before 1961 were removed from the pedigree. About 70% of the animals included in the pedigree had both parents known. Briefly, each genetic group was associated with at least 20 animals. The genetic group -31 that corresponds to the birth year 2010-2019, sex male, and country Poland was associated with most missing sires and assigned to 1,002,069 individuals. The largest number of missing dams was assigned the '-32' group (that is, birth year 2010-2019, sex female and country Poland) and contains 174,954 individuals.

Table 2: Genetic groups definition

Country	Year of birth	Male	Female
	<1960	-99	-99
POL	1960-1969	-1	-2
USA/CAN	1960-1969	-3	-4
OTHERS	1960-1969	-5	-6
POL	1970-1979	-7	-8
USA/CAN	1970-1979	-9	-10
OTHERS	1970-1979	-11	-12
POL	1980-1989	-13	-14
USA/CAN	1980-1989	-15	-16
OTHERS	1980-1989	-17	-18
POL	1990-1999	-19	-20
USA/CAN	1990-1999	-21	-22
OTHERS	1990-1999	-23	-24
POL	2000-2009	-25	-26
USA/CAN	2000-2009	-27	-28
OTHERS	2000-2009	-29	-30
POL	2010-2019	-31	-32
USA/CAN	2010-2019	-33	-34
OTHERS	2010-2019	-35	-36
POL	2020-present	-37	-38
USA/CAN	2020-present	-39	-40
OTHERS	2020-present	-41	-42

The following single-step random regression SNP-BLUP test-day model (Liu et al., 2004) was applied:

$$y = Xh + Wf + Vp + Vu + e,$$

where **y** contains cow's test day fat yield records from the first three lactation, **h** is a vector of fixed effects of herd-test-date-parity-milking frequency, **f** is a vector of fixed lactation curve coefficients which was modelled by the Wilmink function (Liu et al., 2004), **p** is a vector of permanent environmental effects expressed as random regression coefficient coefficients of the Legendre polynomial, **u** is a random additive genetic effects also described by the random regression coefficients of the Legendre polynomials.

GEBVtest method was chosen to perform the validation (Mäntysaari et al., 2010). It involves the preparation of two data sets, a full data set that includes all phenotypic data, and a

truncated data set that corresponds to the whole dataset with the latest 4 years of phenotypic data removed. Validation bulls were defined as bulls with daughters associated with records in the whole dataset but none in the truncated datasets.

The validation bulls were selected based on the full data set based on the following criteria: born between 2015-2019, have over 20 daughters with records.

Validation results was prepared for three lactation and total EBV, which includes:

$$Total\ EBV = 0.5 * 1st\ lactation\ EBV + 0.3 * 2nd\ lactation\ EBV + 0.2 * 3rd\ lactation\ EBV$$

Analyses were conducted using MiXBLUP 3.0 (Vandenplas et al., 2022)

Results & Discussion

For 815 validation bulls, we prepared validation results. For pedigree BLUP with and without genetic groups, validation resulted in b_1 of 1.03 (Table 3) and 1.01 (Table 4), respectively. Using the single-step random regression SNP-BLUP test-day model without genetic groups resulted in b_1 equal to 0.82 (Table 5). After defining the genetic groups, b_1 increased to 0.92 (Table 6). Finally, considering metafounders in ssSNP-BLUP improved the validation performance that achieved a $b_1$1.05.

Adding genotype information and using the single-step random regression SNP-BLUP test-day model resulted in a decreased b_1 of the validation. However, adding genetic groups and metafounders led to an increase of b_1 (Figure 1). Expressed by the R^2 value and correlation between GEBVs from the whole and truncated data sets, the same growing trend can be observed (Figure 2, Figure 3). For Pedigree BLUP without and with genetic groups and single-step random regression SNP-BLUP test-day model without genetic groups, the values of R^2 and correlation are similar, 0.43, 0.46, 0.45 respectively for R^2 and 0.66, 0.70, 0.67 for correlation. They changed when genetic groups and then metafounders were included in the pedigree. The best results were obtained for the scenario with metafounders, yielding R^2 of 0.73 and correlation of 0.86, while for the scenario with genetic groups the R^2 is 0.62 and correlation is 0.76.

Table 3: Results of validation for pedigree BLUP without genetic groups.

Bulls	b_0[1]	b_1[2]	R^2[3]	corr.[4]
1st lactation	-21.928	0.984	0.420	0.648
2nd lactation	-26.448	1.045	0.444	0.667
3rd lactation	-31.238	1.092	0.448	0.670
Total EBV	-8.471	1.030	0.435	0.660

Table 4: Results of validation for pedigree BLUP with genetic groups.

Bulls	b_0	b_1	R^2	corr.
1st lactation	-23.766	0.969	0.473	0.688
2nd lactation	-29.896	1.024	0.492	0.702
3rd lactation	-33.439	1.060	0.492	0.701
Total EBV	-9.304	1.009	0.485	0.696

Table 5: Results of validation for single-step random regression SNP-BLUP without genetic groups

Bulls	b_0	b_1	R^2	corr.
1st lactation	-30.391	0.818	0.441	0.664
2nd lactation	-30.428	0.809	0.457	0.676
3rd lactation	-27.911	0.823	0.467	0.684
Total EBV	-9.990	0.815	0.450	0.670

Table 6: Results of validation for single-step random regression SNP-BLUP with genetic groups

Bulls	b_0	b_1	R^2	corr.
1st lactation	-19.690	0.934	0.621	0.788
2nd lactation	-18.783	0.907	0.613	0.783
3rd lactation	16.873	0.907	0.614	0.784
Total EBV	-6.285	0.919	0.617	0.785

Table 7: Results of validation for single-step random regression SNP-BLUP with metafounders.

Bulls	b_0	b_1	R^2	corr.
1st lactation	-19.655	1.005	0.708	0.841
2nd lactation	-30.137	1.067	0.750	0.866
3rd lactation	-35.647	1.098	0.756	0.869
Total EBV	-8.774	1.046	0.733	0.856

[1] b_0 - intercept

[2] b_1 - slope

[3] R^2 - coefficient of determination

[4] corr. - correlation

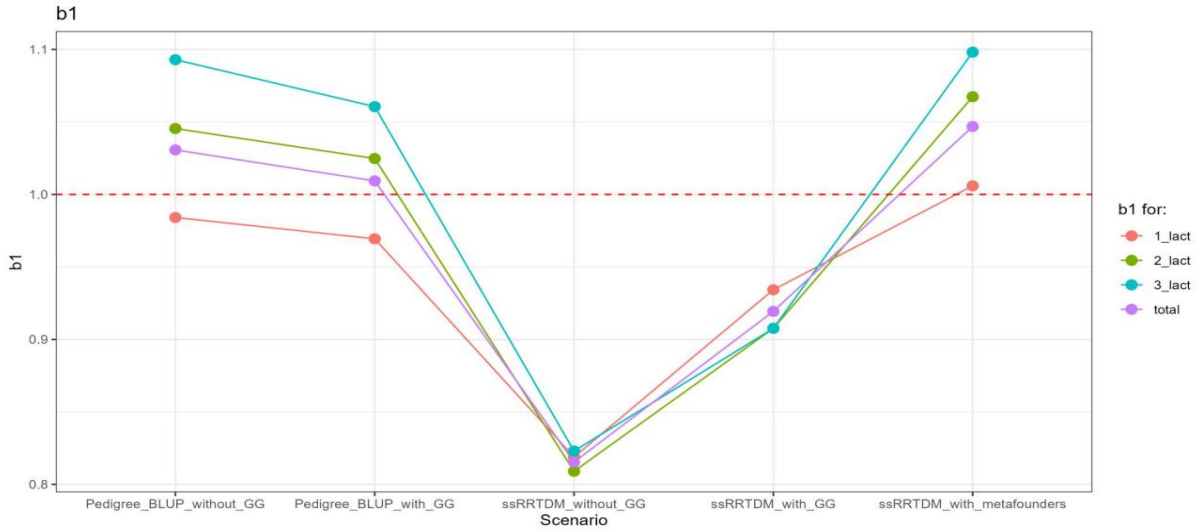

Figure 1. Validation regression coefficient (b_1) for different scenarios.

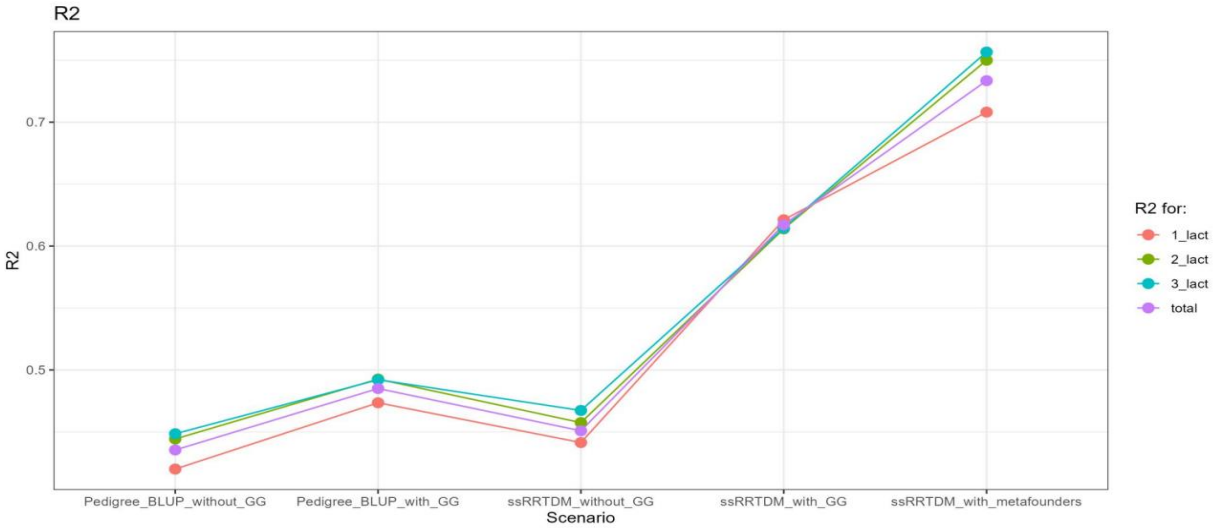

Figure 2. R^2 for different scenarios.

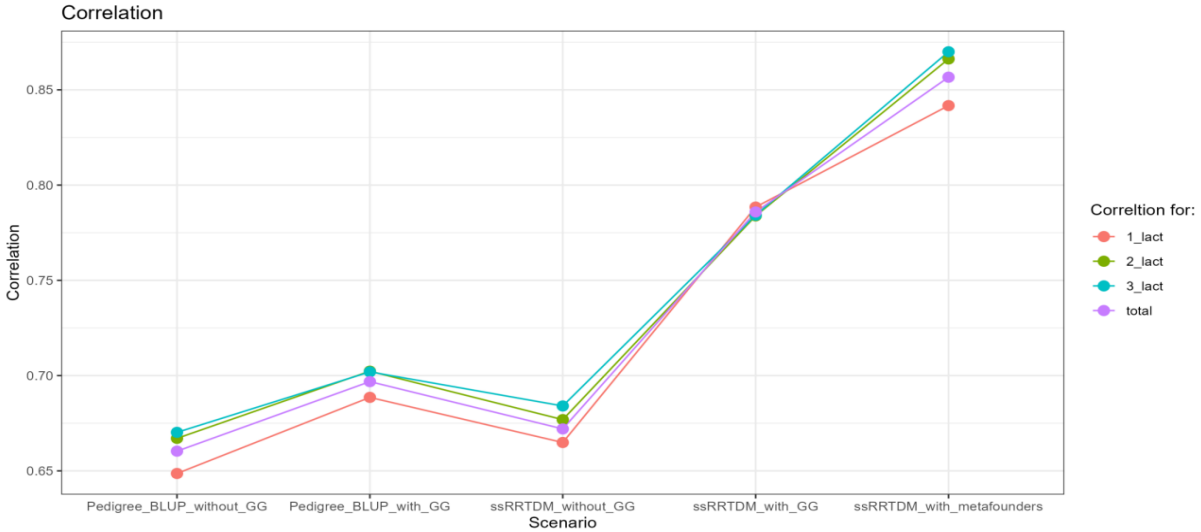

Figure 3. Pearson correlations between (genomic) estimated breeding value obtained from the whole and truncated datasets for different scenarios.

Results & Discussion

Models fitting a pedigree without genetic groups achieved much faster convergence with less iterations, and, obviously, the pedigree BLUP model converged faster than the SNP-BLUP model when fitting the same pedigree. The difference between both pedigree BLUP scenarios is large. The scenario with genetic groups needed 83 minutes more and 2,007 iterations more to get convergence. Similar situations were observed for the three single-step scenarios. The scenario with genetic groups needed 210 minutes more and 2,353 iterations more to get convergence than scenario without genetic groups, while the scenario with metafounders resulted in an intermediate number of iterations and thus the elapsed time (Table 8).

Table 8: Time and iteration per scenario

Scenario	Wall clock time (min)	Number of iterations
Pedigree BLUP without genetic groups	55	273
Pedigree BLUP with genetic groups	137	2280
ssRRTDM SNP-BLUP without genetic groups	154	949
ssRRTDM SNP-BLUP with genetic groups	372	3302
ssRRTDM SNP-BLUP with metafounders	283	2496

Conclusions

The use of alternatives to missing parents in the form of genetic groups or metafounders markedly improves the validation results. Particular improvements are seen in the single-step random regression SNP-BLUP test-day model, where the use of genetic groups first and then metafounders improved the b_1, yielded a model with the higher R2, and achieved higher correlation between GEBVS obtained from the whole and truncated datasets of validation bulls. The reason for this improvement may be the large amount of missing pedigree data for individuals born between 2010 and 2019, so the use of genetic groups and metafounders complements the missing information. The downside of using a more sophisticated pedigree architecture is the increased number of iterations and elapsed time until convergence.

References

Bradford, H., Masuda, Y., Vanraden, P., Legarra, A., and Misztal, I. 2019. Modeling missing pedigree in single-step genomic BLUP. *J. Dairy Sci.*, 102(3), 2336–2346. https://doi.org/10.3168/jds.2018-15434

Legarra, A., Bertrand, J., Strabel, T., Sapp, R., Sánchez, J., and Misztal, I. 2007. Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics*, 124(5),286–295. https://doi.org/10.1111/j.1439-0388.2007.00671.x

Liu, Z., Reinhardt, F., Bünger, A., and 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(6), 1896–1907. https://doi.org/10.3168/jds.s0022-0302(04)73348-2

Mäntysaari, E., Liu, Z., and VanRaden, P. 2010. Interbull validation test for genomic evaluations. *Interbull Bulletin*, 41, 17. https://journal.interbull.org/index.php/ib/article/download/1134/1125

Vandenplas, J., Veerkamp, R., Calus, M., Lidauer, M., Strandén, I., Taskinen, M., Schrauf, M., and Napel, J. T. 2022. 358. MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. https://doi.org/10.3920/978-90-8686-940-4_358

Westell, R., Quaas, R., and Van Vleck, L. 1988. Genetic Groups in an Animal Model. *J. Dairy Sci.*, 71(5), 1310–1318. https://doi.org/10.3168/jds.s0022-0302(88)79688-5