Cross-Validation Assessment of Random Regression Specifications in a Single-Step Genomic Model for Dry Matter Intake

M.F. Schrauf [1*], R.F. Veerkamp [1], R. Bonifazi [1], C.M. Orrett [2], G. de Jong [3], B. Gredler-Grandl [1]

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

[2]CRV B.V., P.O. Box 454, 6800AL Arnhem. The Netherlands.

[3]Coöperatie Koninklijke CRV U.A., P.O. Box 454, 6800AL Arnhem. The Netherlands

** Corresponding author: matias.schrauf@wur.nl*

Abstract

Selection on feed efficiency traits can help to reduce costs and improve sustainability in the dairy cattle industry. Recent advances propose to use a random regression to derive breeding values for dry matter intake (DMI) from longitudinal models. In this study, we conduct a forward cross-validation of different random regression specifications of an animal model for DMI. The specifications combine basis functions for regression over days in milk with varying numbers of factors used in variance component estimation via a factor-analytic approach. Data from 10,766 predominantly Dutch and Belgian Holstein cows, comprising 21,008 lactations and 1,026,192 DMI records from 10 farms, were analyzed. Estimates obtained from partial data (pre-2020) were compared to those from the full dataset (up to early 2024). Multiple sets of focal individuals were used to estimate prediction errors for the models, decomposing global error summaries into intercept bias, slope bias, and correlation; for early, middle, and late lactation stages. The validation results identify random regression specifications that outperform the accuracy of a conventional repeatability model for DMI, in particular on the early and middle stages of lactation. This provides valuable insights for genomic prediction modeling of feed efficiency in cattle.

Key words: cross-validation, random regression, dry matter intake, dairy cattle

Introduction

Feed efficiency is an important trait in dairy cattle breeding due to its significant economic and environmental implications. Until at least 2017, the genetic trend for feed efficiency was slightly negative (Pryce and Bell, 2017; de Jong et al., 2019), primarily due to increased body size and associated maintenance feed requirements (de Jong et al., 2019). While debate continues on whether to include dry matter intake (DMI) directly in breeding goals or consider it through traits like residual feed intake (RFI) (Veerkamp et al., 2013), individual DMI recording remains essential for genetic improvement of feed efficiency in dairy cattle.

Direct measurement of DMI is expensive and logistically challenging (Berry et al., 2014). Genomic prediction models help to address this issue by enhancing the value of phenotypes recorded from genotyped cows. Furthermore, single-step genomic models enable efficient use of information from both genotyped and non-genotyped animals connected through pedigree. Despite these advantages, genetic models for DMI have typically been focused on repeatability models due to limited data availability.

As more DMI records accumulate, there is an opportunity to explore more complex random regression models (RRMs) that can account for changes in genetic effects across lactation. These models provide dynamic predictions of breeding values depending on lactation stage and can be used in conjunction

with predictions for energy sinks and sources (milk production and liveweight changes) to estimate recently proposed traits such as genomic residual feed intake (gRFI; Islam et al., 2020). Even when obtaining a gRFI is not the objective, RRMs can improve the utilization of records from cows at different lactation stages by accommodating higher correlations between records taken close together in time, while allowing for lower correlations between early and late lactation and across lactations (Veerkamp et al., 2013). In contrast, repeatability models assume a genetic correlation of unity and can over-estimate the amount of information for cows with sparsely recorded data.

In this study, we assess the efficacy of these random regression models and compare them to each other with a forward cross-validation technique. This process involves comparing estimates from partial datasets to those from complete datasets, thereby evaluating the predictive accuracy of the models. Through this empirical validation, the study aims to identify the most accurate random regression specifications for estimating breeding values for DMI, offering insights for genomic prediction models in feed efficiency.

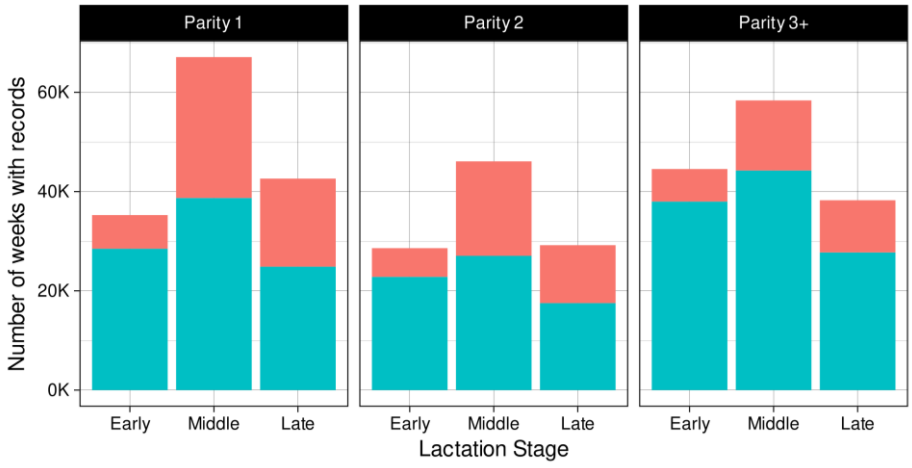
Materials and Methods

Feed Intake Data and Genetic Model

Data used in this study were routinely collected for the analysis of feed intake in dairy cattle in the Netherlands and northern Belgium. Individual feed intake was recorded at 10 facilities: the Dairy Campus of Wageningen Livestock Research, Schothorst Feed Research, ILVO Research Institute, feed companies Trouw Nutrition and AVEVE, and five commercial farms associated with CRV.

Cows with fewer than 3 DMI records and those with less than 50% Holstein breed composition were excluded. Records below 8 kg/day and above 55 kg/day were removed. The final dataset included 10,766 cows, 21,008 lactations, and 1,026,192 DMI records. Daily records were aggregated into 389,967 weekly averages (Figure 1).

Figure 1: Number of weeks with DMI records per parity and stages of lactation. In cyan, measurements recorded before 2020, in red measurements from 2020 onwards. Lactation divided into early (0-9 weeks), middle (10-24 weeks), and late (25+ weeks).



The following statistical model, based on research by Veerkamp et al. (2014), was used to calculate breeding values for DMI:

y = PAR + HM + HY + AGE + LS + B + PERM + A + Res

where:
y: Individual DMI (weekly average)
PAR: Parity, 3 levels (parity 1, 2, and 3+)
EXP: Experiment, a combination of farm and management/experiment effect
HM: Herd*month of calving
HY: Herd*year of calving
AGE: Age at calving per parity, quadratic polynomial
LS: Lactation stage (Days in milk), 4th order polynomial
B: Breed % of the second breed, intercept and slope
PERM: Permanent environment of animal
A: Breeding value of animal
Res: Residual

The random effects PERM and A were specified with random regression models, as described in the next section. For the animal effect A, the numerator relationship matrix was used for pedigree-based models (Henderson, 1976). For single-step genomic models, marker information was integrated following the method of Liu et al. (2014), using a ssSNPBLUP model fitted with the hpblup solver in MiXBLUP (Vandenplas et al., 2022).

Random Regression Specifications

Four random regression model structures were evaluated on days in milk:
1. Repeatability
2. Piecewise-constant
3. Linear
4. Cubic

The repeatability model structure is the simplest, similar to those currently used in genetic evaluations, with a single breeding value for each parity of the cow. The remaining models allow the breeding values of an individual animal to change over days in milk, within the same lactation. The piecewise-constant model can be considered a multi-trait model, where DMI is divided into six different traits depending on the stage of lactation. The linear and cubic models are typical random regression models, modeling the varying breeding value as a polynomial curve of the corresponding order.

All models can be formulated as random regressions, differing only in the basis functions used:
- The repeatability model uses a single constant basis function.
- The piecewise-constant model uses an indicator function for each stage of lactation.
- The linear and cubic models use Legendre polynomial basis functions of degree 1 and 3, respectively.

Variance components for each model were estimated using ASReml (Gilmour, 2019). Except for the repeatability model, variance component estimation was simplified using a factor-analytic approach, iteratively increasing the number of factors until model likelihood stopped improving.

Cross-validation Scheme

A forward cross-validation scheme was used to assess the predictive accuracy of the models. Data were split into partial (records before 2020) and whole (records up to early 2024) datasets. Breeding values for DMI were predicted from the partial dataset and compared to corresponding breeding values from the whole dataset (similarly to Legarra and Reverter, 2018). Each lactation was divided into early (weeks 5 and 10), middle (weeks 15 and 25), and late (weeks 35 and 45) periods for the validation.

To overcome the limitation of each model being validated against itself, a common reference model (piecewise-constant) was selected, and partial predictions of each model

were compared to the whole predictions of this reference model.

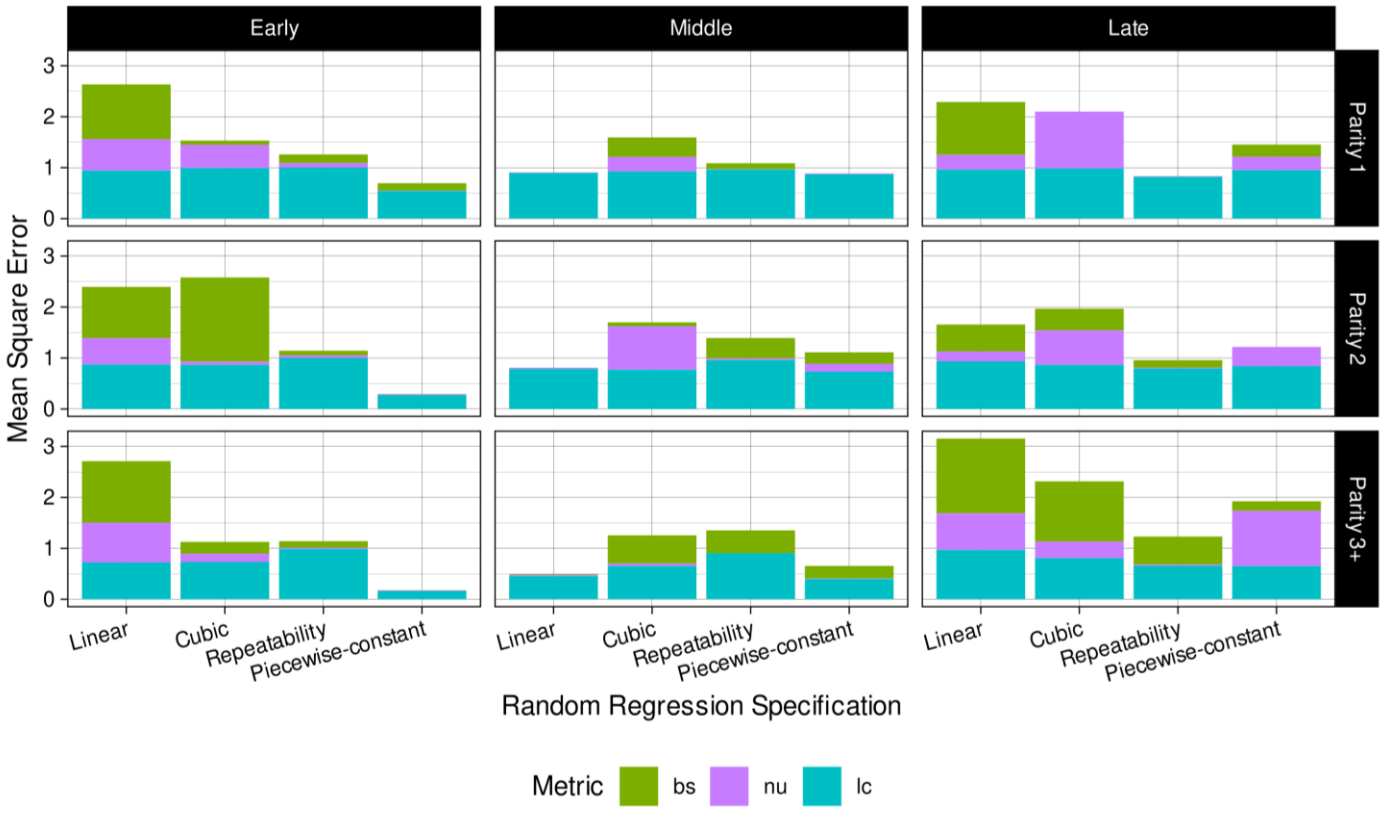
Validation was performed for multiple sets of focal individuals, with metrics reported here for validation cows (the largest focal group). Validation cows were defined as those with at least 3 DMI records after January 1st, 2020, and no DMI records before that date, consisting of 2,958 cows.

Following Gauch et al. (2003), the following validation metrics were calculated (Table 1): Bias Squared (BS), Non-unity of slope (NU), Lack of Correlation (LC), and their sum which equals the Mean Squared Error (MSE).

Table 1: Validation metrics used in this study. Where, 'a' are breeding values for DMI, subindices 'p' and 'w' indicate the partial and whole dataset, respectively.

Validation metric	Quantifies	Related Variable	Formula
Bias Squared (BS)	Level bias	Intercept (b_0)	$(\bar{a}_w - \bar{a}_p)^2$
Non-unity of Slope (NU)	Inflation/deflation	Slope (b_1)	$(1-b_1)^2 Var(a_p)$
Lack of Correlation (LC)	Accuracy	Correlation (r)	$(1-r^2)Var(a_w)$
Means Square Error (MSE)	All discrepancies	BS + NU + LC	$\Sigma_i (a_w(i) - a_p(i))^2$

Figure 2: Validation metrics for different random regression specifications with pedigree-based models. Lactation divided into early (0-9 weeks), middle (10-24 weeks), and late (25+ weeks).



Results and Discussion

Comparison of the four random regression models for predicting DMI breeding values, using pedigree-based models and the piecewise-constant model as a reference, revealed that the piecewise-constant model showed the best overall performance (Figure 2). The piecewise-constant model had the lowest MSE (0.48) and highest correlation (0.41) between predicted and observed values.

The linear model performed relatively well in mid-lactation but poorly in early and late lactation periods. Most of the MSE in early and late lactations was due to intercept and slope bias rather than lack of correlation, suggesting that true genetic effects on DMI are non-linear across days in milk, and the linear model lacks the flexibility to capture these changes.

The cubic model showed variable performance across lactation stages, with the highest MSE in early lactation for second parity and the lowest in middle and late lactation for 3+ parities. The general lack of improvement over the linear model suggests that pedigree information alone is insufficient to predict varying breeding values accurately with a cubic regression, given the current data availability.

The repeatability model showed low bias overall but low accuracy for early and middle lactation periods. The low bias indicates that the typical curve for the true genetic effects does not deviate greatly from the constant breeding value assumed in the repeatability model. However, the lack of correlation in early and middle lactation suggests that the absence of distinctions between different stages of lactation in the repeatability model impacts the predictive accuracy.

Prediction was generally most challenging in early lactation, with the piecewise-constant model showing the lowest MSE in this period. This may be due to metabolic changes during the transition period and reduced data availability in early lactation compared to middle lactation. Late lactation was easier to predict, though it is unclear whether this is due to the pattern of data available for validation cows or a more stable metabolic state at this stage.

Inclusion of genomic information in single-step random regressions (for linear and cubic models) improved both stability within models and consistency across models (Table 2). This improvement was more pronounced for the more complex cubic model (65% vs. 30% improvement in stability, 45% vs. 21% in consistency) compared to the linear model. This suggests that the more efficient use of available records with genomic information might allow for effective use of a polynomial model, contrary to observations with pedigree-based models.

Table 2: Correlations for pedigree and single-step genomic models, between estimated breeding values in partial and whole datasets.

		Validation Target	
Model	Regression	Linear	Cubic
Pedigree	Linear	0.39	0.28
	Cubic	0.29	0.26
Single-step	Linear	0.51	0.34
	Cubic	0.42	0.43

Conclusions

This study demonstrates the potential for improving genetic evaluation of DMI in dairy cattle using more flexible random regression models compared to simple repeatability models. The piecewise-constant approach appears promising, though it may be beneficial to further refine the lactation periods over which prediction is constant in this model. With single-step genomic models, a polynomial random regression model may sufficiently model genetic changes throughout lactations.

Future work could explore additional model structures such as splines, evaluate uncertainty in validation metrics, and combine results across different focal groups. As DMI data continues to accumulate, a reassessment of

model comparisons and optimal recording periods can be useful to ensure optimal use of available information for genetic improvement of feed efficiency.

Acknowledgments

The authors would like to thank the participating farms and research institutions for providing the data used in this study. This research was supported by CRV, Wageningen Research and the Dairy Campus innovation fund.

References

Berry, D.P., Coffey, M.P., Pryce, J.E., De Haas, Y., Løvendahl, P., Krattenmacher, N., Crowley, J.J., Wang, Z., Spurlock, D., Weigel, K. and Macdonald, K., 2014. International genetic evaluations for feed intake in dairy cattle through the collation of data from multiple sources. *Journal of dairy science*, 97(6), pp.3894-3905. https://doi.org/10.3168/jds.2013-7548

de Jong, G., de Haas, Y., Veerkamp, R., Schopen, G., Bouwmeester-Vosman, J. and van der Linde, R., 2019. Feed intake genetic evaluation: Progress and an index for saved feed cost. *Interbull Bulletin*, (55), pp.1-4.

Gauch, H.G., Hwang, J.G. and Fick, G.W., 2003. Model evaluation by comparison of model-based predictions and measured values. *Agronomy Journal*, 95(6), pp.1442-1446. https://doi.org/10.2134/agronj2003.1442

Gilmour, A.R., 2019. Average information residual maximum likelihood in practice. *Journal of Animal Breeding and Genetics*, 136(4), pp.262-272. https://doi.org/10.1111/jbg.12398

Henderson, C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, pp.69-83. https://doi.org/10.2307/2529339

Islam, M.S., Jensen, J., Løvendahl, P., Karlskov-Mortensen, P. and Shirali, M., 2020. Bayesian estimation of genetic variance and response to selection on linear or ratio traits of feed efficiency in dairy cattle. *Journal of dairy science*, 103(10), pp.9150-9166. https://doi.org/10.3168/jds.2019-17137

Legarra, A. and Reverter, A., 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, 50, pp.1-18. https://doi.org/10.1186/s12711-018-0426-6

Liu, Z., Goddard, M.E., Reinhardt, F. and Reents, R., 2014. A single-step genomic model with direct estimation of marker effects. *Journal of Dairy Science*, 97(9), pp.5833-5850. https://doi.org/10.3168/jds.2014-7924

Pryce, J.E. and Bell, M.J., 2017. The impact of genetic selection on greenhouse-gas emissions in Australian dairy cattle. *Animal Production Science*, 57(7), pp.1451-1456. https://doi.org/10.1071/AN16510

Vandenplas, J., Veerkamp, R.F., Calus, M.P.L., Lidauer, M.H., Strandén, I., Taskinen, M., Schrauf, M.F.S.G. and ten Napel, J., 2022, December. MiXBLUP 3.0–software for large genomic evaluations in animal breeding programs. *In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP) (pp. 1498-1501). Wageningen Academic Publishers*. https://doi.org/10.3920/978-90-8686-940-4_358

Veerkamp, R. F., Pryce, J. E., Spurlock, D., Berry, D., Coffey, M., Løvendahl, P., van der Linde, R., Bryant, J., Miglior, F., Wang, Z., Winters, M., Krattenmacher, N., Charfeddine, N., Pedersen, J., and de Haas, Y, 2013. Selection on feed intake or feed efficiency: a position paper from gDMI

breeding goal discussions. *Interbull Bulletin*, (47)

Veerkamp, R.F., Calus, M.P.L., de Jong, G., van der Linde, C., and De Haas, Y. 2014. Breeding Value for Dry Matter Intake for Dutch Bulls based on DGV for DMI and BV for Predictors. *In Proceedings of 10th World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver*. https://edepot.wur.nl/359020