Computers and Electronics in Agriculture 184 (2021) 106106

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Mining data from milk mid-infrared spectroscopy and animal characteristics to improve the prediction of dairy cow's liveweight using feature selection algorithms based on partial least squares and Elastic Net regressions

Lei Zhang [a,*], Anthony Tedde [a,f], Phuong Ho [b], Clément Grelet [c], Frédéric Dehareng [c], Eric Froidmont [c], Nicolas Gengler [a], Yves Brostaux [a], Dagnachew Hailemariam [d], Jennie Pryce [b,e], Hélène Soyeurt [a]

[a] TERRA Teaching and Research Centre, University of Liège - Gembloux Agro-Bio Tech, 5030 Gembloux, Belgium
[b] Agriculture Victoria Research, Agribio, 5 Ring Road, Bundoora, Victoria 3083, Australia
[c] Valorisation of Agricultural Products Department, Walloon Agricultural Research Centre, 5030 Gembloux, Belgium
[d] Faculty of Agricultural, Life and Environmental Sciences, University of Alberta, Alberta, Canada
[e] La Trobe University, Bundoora, Victoria 3083, Australia
[f] National Funds for Scientific Research, 1000 Brussels, Belgium

ARTICLE INFO

Keywords:
Dairy Cow Liveweight
Mid-Infrared
Feature Selection
Bias and Robustness
Modeling

ABSTRACT

Body weight (BW) of dairy cows is relevant for breeding programs and farm management to assess the maintenance requirements, reproduction performance, or health status of cow. Currently, it is still difficult to follow BW changes of individual cows routinely in large herds. Combined with animal characteristics, milk mid-infrared (MIR) spectrum was proposed as an additional source of information to predict BW under the framework of dairy herd improvement (DHI) programs. However, the presence of less informative variables in the prediction equation could impact negatively its robustness. This research aims to improve the robustness of BW regression models by applying a feature selection before modeling. A total of 5,920 BW records composed of animal characteristics and milk MIR spectrum were collected from Holstein cows. Three feature selection algorithms were applied to select the most informative variables: partial least squares regression (PLS) combined with sum of ranking difference (PLS-SRD), PLS combined with uninformative variables elimination (PLS-UVE), and the output of Elastic Net regression (EN). Four herd independent validation sets and the corresponding remained calibration datasets having on average 163 and 1,708 records, respectively, were used to develop models using PLS or EN approaches. Ten-fold cross-validation was conducted to parametrize each model. Parity, days in milk (DIM), milk yield (MY), and two MIR spectral points were selected as relevant variables to predict BW. PLS (root mean square error of validation, RMSEp = 60 kg) and EN (RMSEp = 60 kg) regressions employing these 5 predictors were more robust than the models developed without MIR or using MIR without feature selection. The EN models had a cross-validation root mean square error of around 53 kg. The 2 MIR points explained up to 4.20% variation in predicting BW. The RMSE of validation sets using another brand of spectrometer were around 64 kg. This study confirms the possibility to predict an indicator of BW from animal characteristics and MIR variables. The variable selection procedures improved the model's robustness and transferability. The accuracy

Abbreviations: **MIR**, mid-infrared; **DHI**, dairy herd improvement; **BW**, Live body weight; **PLS**, Partial least squares regression; **SRD**, Sum of ranking difference; **UVE**, Uninformative variables elimination; **EN**, Elastic Net regression; **PLS-SRD**, Partial least squares regression (PLS) combined with sum of ranking difference; **PLS-UVE**, PLS combined with uninformative variables elimination; **PLS-NoMIR**, PLS regression without MIR spectrum; **PLS-All**, PLS regression using animal characteristics and MIR spectrum; **PLS-FS**, PLS regression using variables based on feature selection; **EN-FS**, EN regression using variables based on feature selection; **DIM**, Days in milk; **MY**, Milk yield; **RMSE**, Root mean square error; **BCS**, Body condition score; **GH**, Global H distance; **PC**, Principal components; **LV**, Latent variables; **RPD**, Ratio of performance to deviation; **SD**, Standard deviation; **C value**, Reliability of predictor; **RRM**, Rank reliability measure; **ICAR**, International Committee for Animal Recording.

* Corresponding author at: Avenue de la faculté d'agronomie, 8, B5030 Gembloux, Belgium.
E-mail address: lei.zhang@doct.uliege.be (L. Zhang).

https://doi.org/10.1016/j.compag.2021.106106
Received 19 August 2020; Received in revised form 15 February 2021; Accepted 13 March 2021
Available online 25 March 2021

of BW prediction seems to be sufficient to provide useful information for breeding program and farm management decisions under a DHI framework.

1. Introduction

Phenotypes related directly or indirectly to feed efficiency have to be considered in animal breeding and farm management (Miglior et al., 2017; Zhang et al., 2020). Quantifying the cow feed intake is important in order to accurately compare feed efficiency between dairy herds. Unfortunately, measuring individual feed intake is difficult and expensive. However, the body weight (BW), which is one of the indicators of importance to monitor the energy balance which is related to the cow health, reproduction, and feed efficiency (Thorup et al., 2012; Zhang et al., 2020), can be used to approximate maintenance requirements (Haile-Mariam et al., 2014). For breeding purposes, a limited number of BW measurements can be sufficient. However, for management purpose, the estimation of BW changes on a routine basis is more relevant and this requires repeated estimates (Mäntysaari and Mäntysaari, 2015).

For decades, strategies for the use of BW as a management tool have been explored in both production and research conditions (Bewley and Schutz, 2008; Maltz et al., 1997). Moreover, assessing individual information like BW is required to improve the accuracy of genetic parameters estimated for this trait (Korver, 1988). Thus, efforts have been made to monitor dairy cow BW at large and individual levels (Song et al., 2018; Soyeurt et al., 2019). Several methods exist to quantify or estimate the BW of dairy cows such as the electronic weighing scale, the 3-D camera (error around 43 kg) (Kuzuhara et al., 2015), the prediction of BW from body condition score (BCS) (error around 47 kg) (Haile-Mariam et al., 2014), the BW prediction using animal conformation traits (error varied between 37 kg and 110 kg) (Banos and Coffey, 2012; Haile-Mariam et al., 2014; Vanrobays et al., 2015). However, the first and second methods are not always available on commercial farms due to their cost. The third and fourth ones are laborious and time-consuming (Soyeurt et al., 2019; Yan et al., 2009) and subject to the assessor's ability (Edmonson et al., 1989). Moreover, the last one is often made only one time during the cow productive life in most countries (Soyeurt et al., 2019). However, dairy herd improvement (DHI) programs can provide beneficial information to estimate BW at individual and large level. For instance, BW can be predicted from the milk yield, number of lactation and parity (Enevoldsen and Kristensen, 1997; Kuzuhara et al., 2015; Song et al., 2018). Moreover, a larger part of the individual variability of BW can be assessed by adding the test-day milk mid-infrared (MIR) spectrum (Soyeurt et al., 2019) in the predictive model.

However, some issues could exist which indicate that these preliminary BW prediction equations could be improved. First, expanding the size of the calibration set would be meaningful to cover a larger variability of BW. Second, because the accuracy and the robustness of a predictive model evaluated from the same calibration set partly depend on the predictors used (Mehmood et al., 2020), the elimination of uninformative variables from the BW model should be relevant to avoid over-fitting.

Therefore, the accuracy and robustness of a calibration model could be improved by optimizing the variable selection before modeling (Cai et al., 2008). To address this issue, several informative variable selection algorithms have been developed such as the sum of ranking differences (SRD) procedure (Kalivas et al., 2015; Tencate et al., 2016), the uninformative variables elimination (UVE) (Centner et al., 1996), and the use of the output of Elastic Net (EN) penalized regressions. In brief, compared to the most usual way to parametrize a model that is only based on the root mean square error of cross-validation (RMSEcv), SRD procedure selects the most relevant predictors based on the computation of considered merits derived from statistical parameters denoting the model complexity, variance and bias (e.g., RMSEcv, reliability of regression coefficients, L2 norms). This is conducted without being affected by parameters' weight allocation problems (Nie et al., 2019). The UVE combined with Partial Least Squares (PLS) is another method allowing to select the most relevant subset of predictors. The hypothesis of this method is that the most informative variables have a higher importance in the regression compared to the artificial noisy variables. The EN method was proposed to fix the problems in the presence of high correlations among variables; it select groups of correlated variables and conducts continuous shrinkage, and it can also simultaneously conduct variable selection; the output of this algorithm can be considered as selected variables (Zou and Hastie, 2005).

The objective of this research was to improve the robustness of BW prediction models, based on cow's characteristics and milk MIR spectrum, by selecting the most informative variables based on the feature selection algorithms mentioned above.

2. Materials and methods

2.1. Training dataset

For the data from herd1 to herd11 (Fig. 1), a total of 1,915 records were collected from 363 Holstein cows during 2007 to 2016 from the herds of following institutions: University of Liège (Liège, Belgium), Walloon Agricultural Research Centre (Gembloux, Belgium), University of Alberta (Alberta, Canada), Aarhus University (Tjele, Denmark), Agri-Food and Biosciences Institute (Northern Ireland), University College Dublin (Dublin, Ireland), Walloon Breeding Association (Ciney, Belgium), Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany). Part of those data (N = 754) was initially used by Soyeurt et al., (2019) and part of those data were from Genotype Plus Environment (GplusE) project (http://www.gpluse.eu). Parity ranged from 1 to 11 and the number of days in milk (DIM) ranged from 1 to 512 days. Milk yield (MY) was recorded on daily basis and fat and protein contents were measured locally using Foss MilkoScan FT+ or FT6000 spectrometers (Foss, Hillerod, Denmark) or Standard Lactoscope FT-MIR automatic (PerkinElmer, Waltham, United States) in Walloon Agricultural Research Center (Gembloux, Belgium). The spectrum provided by the milk mid-infrared analysis was also recorded and standardized according to the procedure proposed by Grelet et al. (2017). Then, a first-derivation was applied to these data with a window of 5 spectral points to correct the baseline drift. The weighing scales used to measure the BW were manufactured by Fullwood (Shropshire, UK), Gallagher (Canley, UK) or Griffiths Elder (Bury St Edmunds, UK) instruments. The dataset was edited using the following editing. First, the records were edited according to the International Committee for Animal Recording (ICAR) standard (ICAR, 2017): milk yield between 3.0 and 99.9 kg/d, fat content between 1.5 g/dL and 9.0 g/dL of milk, and protein content between 1.0 g/dL and 7.0 g/dL of milk. Before modeling, the DIM records were split into 35 classes of 15 days. Parity records were split into five classes (i.e., 1, 2, 3, 4, and 5 or more). After removing the noisy MIR regions, which are well known to be related to water absorbance as described by Soyeurt et al. (2019), the first derived absorbance values located in the range of 950–1,600 cm^{-1}, 1,750–1,800 cm^{-1}, and 2,600–3,000 cm^{-1} (i.e., 277 spectral variables) were merged to DIM classes, parity classes, MY and BW to constitute the training dataset. So, this contained 1,871 records composed of 280 predictors collected from 355 cows (i.e., 5.27 records per cow on average).

2.2. Prediction models

As presented in Fig. 1, four herd independent validation datasets

(herd5, herd7, herd10 or herd11) having approximately the same number of records in each set were used in this study to assess the performance of the developed models. Consequently, the calibration sets using the remaining samples had also a similar number of records between them. For each calibration set, a principal component analysis was performed on the spectral data to calculate the Global H distance (GH) of each record to detect potential spectral outliers. The number of principal components used explained 99% of the spectral variability. Records with a GH distance higher than 5 were discarded.

A common modeling procedure was applied on each calibration set (Fig. 2). First, the predictors were scaled and centered. Second, a 10-fold cross-validation was used to choose the best parametrization of models (i.e, optimum number of latent variables (LV) for PLS and the best penalty for EN regression). The best parametrization was the one for which the next value for the considered parameters (LV or penalty) did not contribute to a substantial reduction in the RMSEcv. Third, the potential BW outliers were discarded based a residual analysis. If the prediction residual was higher than the mean + 3SD or lower than the mean – 3SD, the BW record was discarded. Fourth, PLS models were rebuilt using this second edited data set and the obtained regression was validated using the corresponding validation set. Finally, the model performance was assessed by calculating the calibration (RMSEc), cross-validation (RMSEcv) and validation RMSE as well as the ratio of performance to deviation (RPD) which was the ratio of the standard deviation of BW variable to the RMSEcv. The relative error was quantified by dividing RMSEp by BW mean calculated from the validation set. The robustness of the models was assessed according to the difference between RMSEcv and RMSEp, since in robust calibration models, the RMSEp is equal or nearly equal to the RMSEcv (Müllertz et al., 2016; Wang et al., 2018).

Four different modeling approaches were tested and compared: PLS regression including only cow's characteristics (parity + MY + DIM) named PLS-NoMIR; PLS regression using cow's characteristics and MIR spectrum (parity + MY + DIM + 277 MIR variables) as done by Soyeurt et al. (2019) and named as PLS-All; PLS and EN regressions employing the variables selected in common by the 3 used feature selection algorithms, named as PLS-FS and EN-FS, respectively. The feature selection algorithms used are explained further in this section. Consequently, a total of 16 models were built to predict test-day BW of dairy cows (Fig. 1). Then, Pearson correlation coefficients between predictions were calculated to assess the prediction consistency between models. The optimal model will be the one having low RMSEp and RMSEcv combined with a small difference between RMSEcv and RMSEp. The performance of this optimal model was also evaluated through the estimation of RMSEp per class of 60 kg of BW. The C value of the absolute regression coefficients was used to assess the contribution of each predictor to the BW prediction.

2.3. Feature selection

Based on the working hypothesis that the most informative variables must be selected by different feature selection algorithms from different training sets, three feature selection algorithms were implemented on the 280 variables included in the 4 different training sets. Then, from those 12 selections, the common selected variables were grouped as they were assumed to be the most informative variables. In the process of

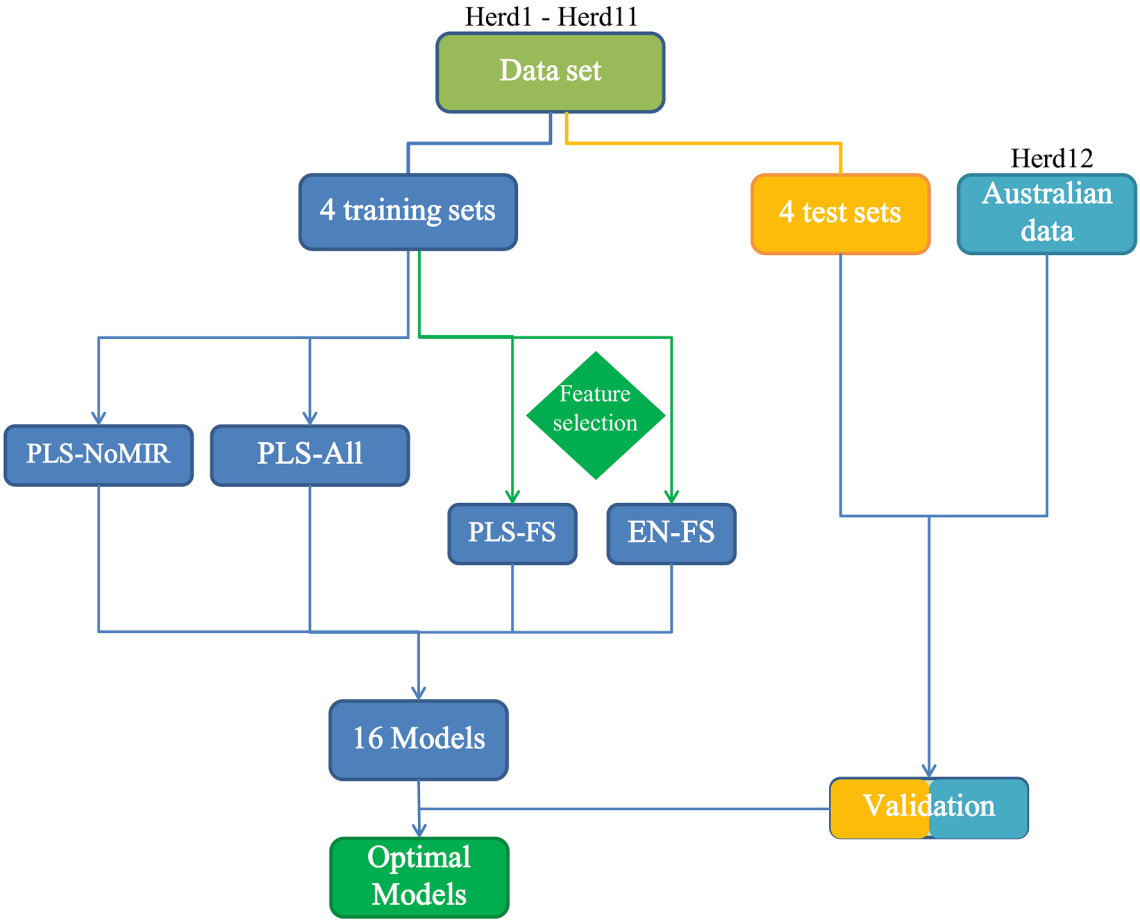

Fig. 1. The technical flow chart of the different modeling approaches used in this study. (*PLS-NoMIR = Partial least squares regression included milk yield, parity, and days in milk; PLS-All = PLS using milk yield, parity, days in milk and 277 MIR data; PLS-FS = PLS using the 5 variables selected by feature selection; EN-FS = Elastic Net regression using the 5 variables selected by feature selection*).

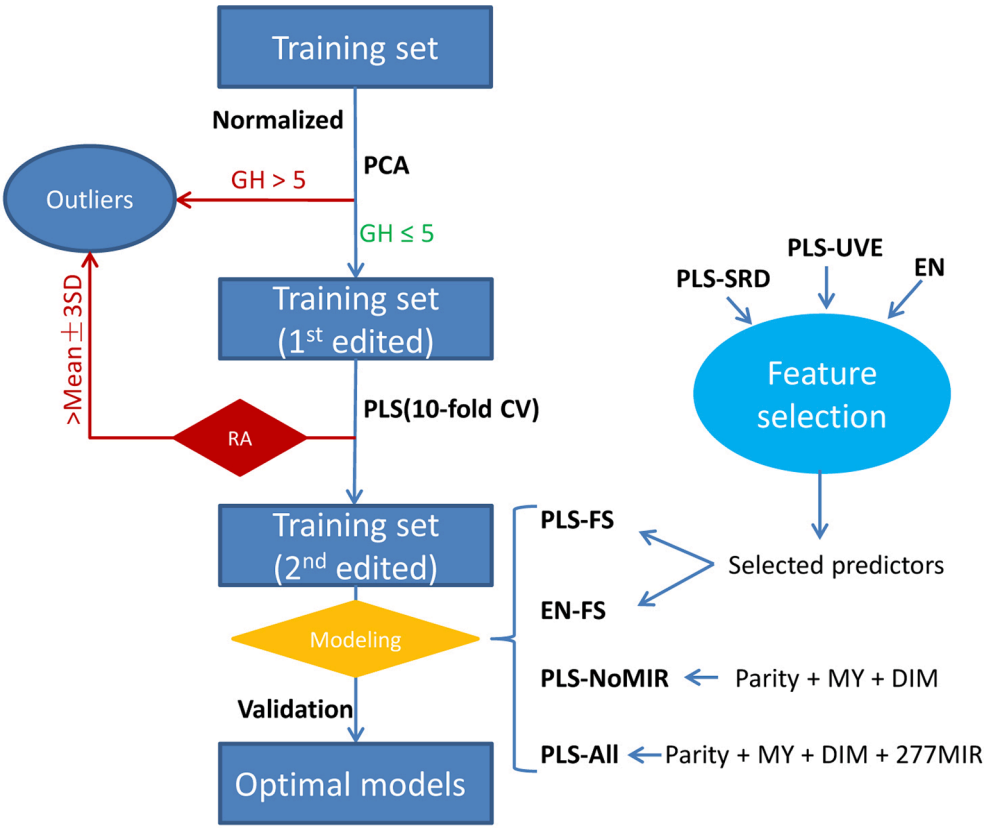

Fig. 2. The scheme of the model specifications and parameter choice for different modeling approaches. (*PCA = Principal component analysis; GH = Global H distance; SD = Standard deviation; RA = Residual analysis; PLS-NoMIR = Partial least squares regression included milk yield, parity, and days in milk; PLS-All = PLS using milk yield, parity, days in milk and 277 MIR data; PLS-FS = PLS using the 5 variables selected by feature selection; EN-FS = Elastic Net regression using the 5 variables selected by feature selection; PLS-SRD = Partial least squares regression combined with sum of ranking difference; PLS-UVE = Partial least squares regression combined with uninformative variable elimination; EN = Elastic Net regression; MY = Milk yield; DIM = Days in milk*).

each modeling algorithm across the feature selection section, 10-fold cross-validation will be used to choose the best parameter. In brief, those 10-fold cross-validations will be conducted in simply the same way (i.e., randomly separate the training set into 10 folds) except based on the different training sets or modeling approaches.

2.3.1. Feature selection based on sum of ranking difference algorithm (SRD)

The first variable selection method was based on SRD algorithm proposed by Kalivas et al. (2015). A 10-fold cross-validation was firstly used to predict BW from the 280 scaled and mean-centered predictors. Consequently, 10 PLS models were built, providing 10 regression coefficients per predictor. For each predictor, the mean and standard deviation (SD) of those coefficients were calculated to compute the reliability (C) (i.e., mean divided by SD (Centner et al., 1996)). The higher the value of C, the more informative and reproducible is the predictor. In a second step, following a decreasing value of C, predictors were included into new PLS models one by one. The tested number of LV was the one considered as optimal in PLS-All and the ones around it having a similar mean and SD of RMSEcv. Then, for each number of LV, different models were tested considered a different number of predictors ranging from LV + 1 to 280. For each tested model, 8 merits were calculated in order to assess the model relevancy. Merit 1 was the jaggedness (J) denoting the model complexity and the degree of over-fitting of a regression model (Eq. (1)) (Gowen et al., 2011).

$$J = \sum_{i=2}^{m} \sqrt{(b_i - b_{i-1})^2} \tag{1}$$

where the b_i denotes the regression coefficient (b) of i^{th} predictor of a model; m is the number of predictors in this model. Merit 2 (called B2) corresponds to the L2 norm of a model (2) and reflects the variance of the model.

$$B2 = \sqrt{\sum_{i=1}^{m} b_i^2} \tag{2}$$

where b_i is the regression coefficient (b) of the i^{th} predictor in a model; m is the number of predictors in this model. Merit 3 was the RMSEcv denoting the model accuracy. Merit 4 (C1, Eq. (3)) and Merit 5 (C2, Eq. (4)) reflect the model bias and variance.

$$C1 = \left(\frac{\|B2\| - \|B2\|min}{\|B2\|\max - \|B2\|min}\right) + \left(\frac{\text{RMSECV} - \text{RMSECV}min}{RMSE\text{CV}\max - RMSE\text{CV}min}\right) \tag{3}$$

$$C2 = \left(\frac{\|J\| - \|J\|min}{\|J\|\max - \|J\|min}\right) + \left(\frac{\text{RMSECV} - \text{RMSECV}min}{RMSE\text{CV}\max - RMSE\text{CV}min}\right) \tag{4}$$

where the max and min represent the maximum and minimum values of B2 or J observed among the models. Merit 6–8 were defined after the calibration process and correspond to the absolute value of $R^2 - 1$ ($|R^2 - 1|$) where R is the calibration correlation coefficient, the absolute value of the regression slope − 1 (|slope − 1|), and the absolute value of the estimated intercept (|intercept|) calculated from the linear regression fitted from the measured and predicted BW records. The lower the values of the 8 merits, the better the PLS model is fitted.

The calculation of SRD value was based on a matrix, which was constituted by 8 rows and up to 822 columns. The workflow is shown in Fig. 3. First, to make the merits comparable at a same amplitude, all merits calculated for each tested model were normalized by models. Second, the gold-rank was defined using the minimum normalized merit value obtained from column (i.e., among the models) and ranked according to an increasing trend by row (i.e., among the merits). Third, for each model (i.e., column), a model ranking based on an increasing trend of their normalized merits value were defined. Fourth, the ranking difference of each model was calculated using the gold-rank and the corresponding rank of each merit of each model. Finally, the sum of ranking

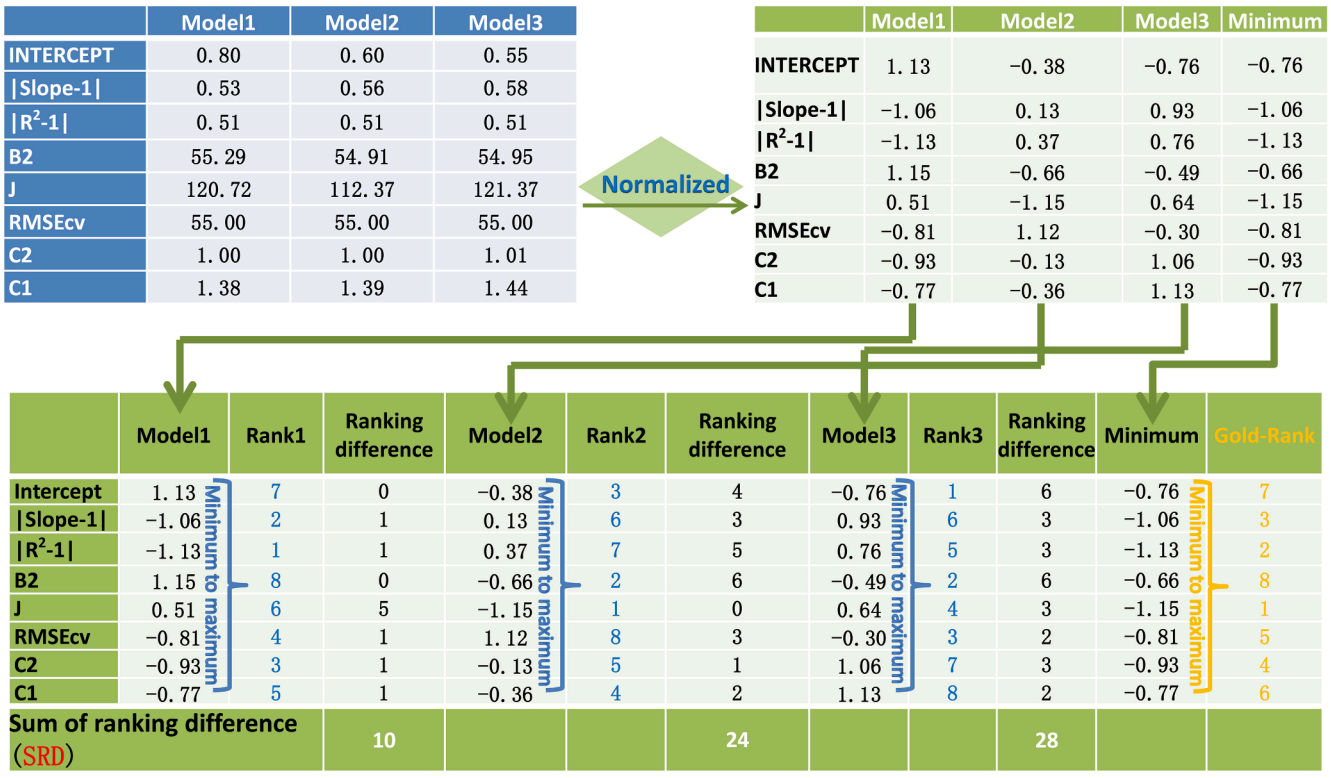

Fig. 3. Workflow of sum of ranking differences (SRD) algorithm.

difference of each model was calculated by accumulating all ranking differences obtained for the considered modeling. Thus, from the example illustrated in Fig. 3, Model 1 has the best prediction performances among these 3 models due to its minimum SRD value. Furthermore, to assess the consistency of the ranking result, the rank reliability measure (RRM) (Eq. (5)) (Tencate et al., 2016) was calculated from the SRD values based on the 8 merits and used to choose the optimal model and the corresponding predictors. The optimal model was the one with the lowest RRM value. For all the tested models, the RRM was calculated as follows:

$$RRM_i = \left(\frac{\sigma_{FR_i} - \sigma_{FR_{min}}}{\sigma_{FR_{max}} - \sigma_{FR_{min}}}\right) + \left(\frac{\overline{FR}_i - \overline{FR}_{min}}{\overline{FR}_{max} - \overline{FR}_{min}}\right) \quad (5)$$

where σ_{FR_i} and $\overline{FR}_i$ are the standard deviation and mean of SRD scores of tested models and scaled to range at 0 and 1 using the minimum and maximum values of all the models. Finally, the predictors included in the optimal model having the lowest RRM value were considered as the most interesting combination of predictors. If equal RRM were obtained for different models, the one with the lowest RMSEcv was considered as having a better fitting.

2.3.2. *Feature selection based on uninformative variable elimination (UVE) algorithm*

The second method used for the selection of informative variables combined PLS and UVE which was proposed by Vitezslav et al. (Centner et al., 1996). First, the predictors were scaled and centered. Second, a matrix of artificial noise variables having the same number of rows and columns than the calibration set was randomly created and added into the predictor matrix column by column. The minimum and maximum value of the noise was set according to the range of the MIR spectra used (–0.35 and 0.36 in this case) in order to have the same magnitude than the MIR signal. A 10-fold cross-validation was applied on the new predictor matrix. A Monte Carlo process (Li et al., 2009) was simulated 8 times during the PLS procedure in order to acquire a stability in the estimation of the regression coefficient from both the MIR and artificial noise variables. Finally, 10 PLS models were generated. Similar to the SRD method, the C value was calculated for each predictor. The information amplitude of the noise must be determined in advance to estimate the suitable cutoff. The amplitude was set at 10^{-11} and an arbitrary value k was set as proposed by Cai et al. (2008). The k value was adopted based on the average of C value estimated for each predictor from the artificial noise matrix. Finally, the predictors having higher C rank than the artificial noises were selected as the most informative variables.

2.3.3. *Feature selection based on Elastic Net regression (EN)*

The third feature selection method was based on the output of the penalized EN regression including the whole scaled and centered predictors. The alpha parameter was set at 0.5 to be between Ridge and LASSO penalized regressions. The tested range for the penalty (called lambda) varied between 0.5 and 1. The optimal lambda was fixed according to the obtained RMSEcv estimated from a 10-fold cross-validation. Finally, the variables with a regression coefficient different than 0 were considered as informative variables.

2.4. *External validation*

To assess the model transferability, the 16 developed regressions were further validated using a completely independent data set coming from another country (i.e., herd12, validation RMSE expressed as RMSEv) and MIR spectral data obtained from another brand of spectrometer. This validation set totalized 4,005 records collected between 2015 and 2017 from 231 cows belonging to the research farm of the Victorian Department of Jobs, Precincts and Regions (Melbourne, Australia). MIR spectra as well as fat and protein contents were provided by a Bentley spectrometer (Model 2000, Chaska, MN, USA). The BW was measured using walkover scales (DeLaval, Tumba, Sweden) as described by HO et al. (Ho et al., 2019). As the BW equation was initially developed from milk Foss MIR spectra, the Bentley spectra were standardized (Grelet et al., 2017) to be expressed on the Foss basis. The RMSE

estimated between the fat and protein contents predicted externally from the standardized spectra and the ones provided by the Bentley spectrometer were 0.11 and 0.27 g/dL of milk, respectively. To ensure the quality of the standardized spectral data, the absolute difference between the fat content predictions (standardized vs. Bentley) was calculated for each record. Records having an absolute fat difference above 0.30 g/dL of milk were discarded. Then, the editing procedure and data editing were the same as the ones used for the training dataset. The final validation dataset contained 3,956 records. The average parity was 2.05 with a range comprised from 1 to 8. The DIM records ranged from 37 to 161 days with a mean of 104 days.

2.5. Computation

All computations and modeling were performed using R software (version 3.6.2) (R Core Team, 2019). The descriptive statistics were analyzed using base package (R Core Team, 2019). The PLS were implemented using the pls package (Bjørn-Helge Mevik, 2019) and the caret package (Max Kuhn et al., 2019) as also for EN regression. The PLS-SRD was computed using a homemade script using R software (version 3.6.2). The PLS-UVE procedure was performed partly from the plsVarSel package (Mehmood et al., 2012)).

3. Results

3.1. Descriptive statistic

The descriptive statistic of the edited training dataset and for the Australian data (herd12) is shown in Table 1. From the training dataset (i.e., herd1 until herd11), the average milk yield was 30.51 ± 11.99 kg/d with a range comprised between 4.30 and 60.50 kg/d. The average fat and protein contents were 3.96 and 3.29 g/dL of milk, respectively. The BW varied between 309 and 984 kg with an average of 619 ± 79 kg. From Australian data (i.e., herd12), the average BW was 550 kg and ranged between 340 kg and 770 kg with a standard deviation of 65 kg.

As shown in Fig. 4, for herds used in the modeling sets (i.e., from herd 1 to herd 11), the mode of studied parity was 2 (Fig. 4A). The DIM records covered the 35 DIM classes (classes over 25 were shown as ≥ 25) and most records were distributed in the 1–4 DIM classes (Fig. 4B). BW records were normally distributed (Fig. 4C) and globally, BW increased with the parity to a stable trend above the third parity (Fig. 4D). The average BW was 540, 607, 662, 666 and 694 kg for first, second, third, fourth and fifth or more parity respectively.

3.2. Predictive models and informative variable selection

Table 2 summarizes the calibration and validation performances observed for the 16 developed models. As R^2 of a model depends highly on the data structure like the distribution and the range of the data (Davies and Fearn, 2006), RMSE is the most relevant statistical parameter to assess the performance of a model (Grelet et al., 2020). So, only validation RMSE are mentioned in Table 2. Discarding of potential BW outliers from residual analysis lead to a slightly different number of records between the calibration sets. Therefore, the 10-fold cross-validation performances cannot be fully compared between models using the same dataset. However, we can compare the validation performances between methods and assess the variability of prediction accuracy for one method thanks to the use of several validation datasets. So, the means and SD of RMSEcv, RMSEp and RMSEv as well as the absolute differences between RMSEp and RMSEcv estimated for each kind of models are mentioned in Table 3.

The averaged RMSEp for PLS-NoMIR was 60 ± 5.57 kg (Table 3). The validation relative prediction error and RPD varied from 8.90% to 13.14%, and 1.4 to 1.5, respectively. The absolute difference between averaged RMSEcv and RMSEp was 8 kg. Higher RMSEp were observed for herd5 and herd10 but more constant error was observed between models using herd12 (RMSEv) (Table2).

The PLS-All models had LV number comprised between 5 and 6. The inclusion of 277 spectral data decreased the reproducibility of the models. Indeed, the SD of RMSEcv for PLS-All was more than 3 times higher than the one observed for PLS-NoMIR (Table 3). The same phenomenon was observed for RMSEv. Only the SD of RMSEp decreased (Table 3). The model calibrated using a training set leaving herd10 and herd11 out had the worse prediction accuracy (Table 2). The RPD ranged from 1.48 to 1.63. The relative error of the RMSEp ranged between 9.04% and 15.46%. The average absolute difference between RMSEcv and RMSEp was 11 kg (Table 4) which was higher than the models from PLS-NoMIR approach. This high difference and the high observed SD of RMSE suggested a low robustness of PLS-All models.

The instability of PLS-All models could be related to the presence of uninformative or non-reproducible in time or among instruments variables in the model. Indeed, the presence of such variables can impact negatively the robustness of the model (Centner et al., 1996). This confirms the relevance of testing different feature selection algorithms to select the best combination of predictors. Table 4 summarizes the results of the feature selections realized in this study.

Multiple models were created during the PLS-SRD selection due to the inclusion of different LV number and predictors ranked based on the reliability of their regression coefficients. As mentioned in the materials and methods section, the number of LVs tested for PLS-SRD feature selection was the one selected for PLS-All and the ones around it having similar mean and SD of RMSEcv. Therefore, the range of tested LV were 5–7, 4–5, 5–6 and 4–5, respectively for the first, second, third and fourth calibration datasets. Based on the number of tested LV and the number of predictors possible for each model (ranged from LV + 1 to 280), 822 models were built from the first calibration set (275 for LV = 5, 274 for LV = 6, and 273 for LV = 7). The number of models tested for the 3 other calibration sets were 551, 549 and 551, respectively (Table 4). Globally,

Table 1
Descriptive statistics of edited data sets.

Herd	N	BW (kg)		Days in milk (days)		Parity		Milk yield (kg/d)		Fat (g/dL of milk)		Protein (g/dL of milk)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	32	680	71.	220.8	142.2	2.66	1.21	21.83	8.07	3.91	0.52	3.53	0.47
2	41	680	79	217.3	112.5	3.00	2.44	28.67	9.44	4.27	0.54	3.54	0.36
3	60	603	38	166.8	101.6	2.10	1.00	21.68	6.28	4.03	0.68	3.27	0.45
4	70	656	79	159.3	85.5	3.54	1.05	21.20	6.73	4.18	0.73	3.34	0.39
5	149	618	72	185.3	112.2	2.56	1.57	19.68	7.04	3.97	1.01	3.55	0.51
6	358	628	69	125.4	77.6	2.36	1.35	19.39	5.20	3.16	1.13	3.11	0.31
7	188	612	84	28.0	13.1	2.80	1.57	30.29	11.07	4.15	0.61	3.19	0.33
8	635	599	86	28.5	12.6	2.40	1.55	37.69	10.47	4.03	0.86	3.39	0.34
9	23	573	51	26.0	11.2	2.48	1.53	34.48	9.36	3.95	0.66	3.15	0.29
10	180	612	69	28.1	12.4	2.44	0.64	41.30	6.66	4.58	1.42	3.18	0.35
11	135	654	57	30.0	12.0	3.78	1.44	34.76	7.37	4.47	0.88	3.10	0.33
12	3,956	550	65	104.0	18.7	2.05	1.06	25.77	4.95	4.18	0.96	3.17	0.34

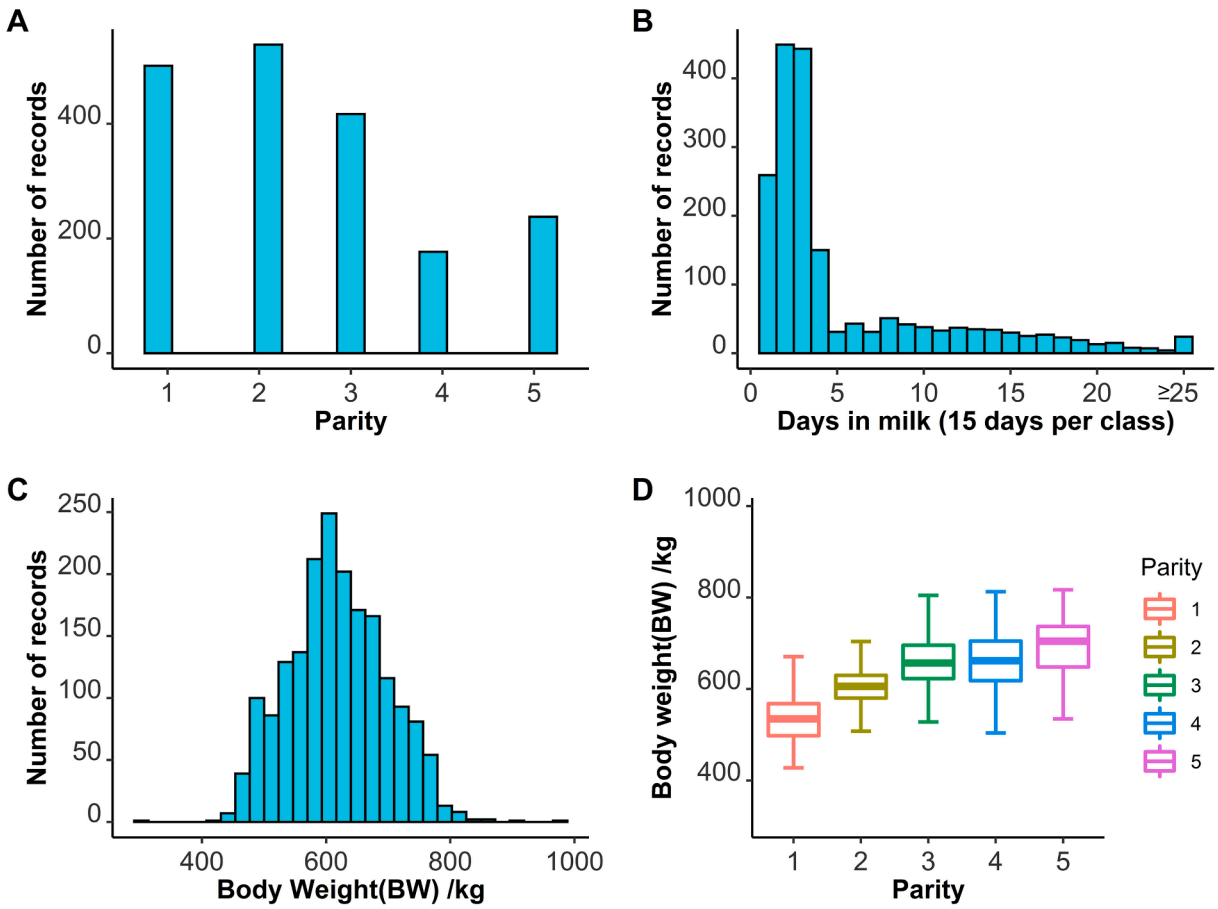

Fig. 4. Distribution of parity (A), days in milk (B), and body weight (C), and the body weight evolution by parity (D) from the edited training data sets.

RRM value obtained for each model fluctuated for all calibration sets along with the change of LV and the number of predictors included in the model (Fig. 5). The range was comprised from 0 to 1.69. The optimal model is the one having the lowest RRM. So, for the first dataset, the 79th, 109th and 121st models had the lowest RRM value (around 0.02; Fig. 5A) revealing a similar robustness. However, 109th model presented the lowest RMSEcv. So, the 114 predictors including in this model into 5 LVs were considered as the most informative based on the first calibration set (i.e., leaving herd5 out). The same procedure was repeated for all other datasets and the results were summarized in Table 4. The best models are illustrated for each dataset in Fig. 5 by the red arrow. From PLS-SRD results, the number of selected variables ranged from 61 to 138 predictors (Table 4) with an average of 105 predictors. In other words, the decrease of features number varied between 51% and 78% of the whole predictors.

The PLS-UVE method selected 106, 120, 69 and 116 informative variables based on the 4 datasets used (Table 4). As for the PLS-SRD method, a strongly lower number of predictors was selected using the third dataset. Except for the third and fourth datasets, PLS-UVE selected a lower number of predictors compared to PLS-SRD. The number of LVs considered in PLS-UVE models were similar to the ones used by PLS-SRD models and turned around 4–5.

Using EN feature selection, the number of selected variables ranged from 62 to 73 (Table 4). The decrease of variable number was stronger here compared to PLS-SRD and PLS-UVE except for the third dataset. The number of discarded samples after the residual analysis was around 27 records which was similar to the ones obtained from PLS-SRD and PLS-UVE methods. The lambda penalty fixed based on the 10-fold RMSE was the same for all datasets and set to 0.5.

Based on the hypothesis that the most informative variables must be selected by different feature selection algorithms from different training sets, the intersection of selected features was defined (Fig. 6). Therefore, from those 12 selections (i.e., 4 training sets and 3 feature selection algorithms), only 5 predictors were in common. They were then included into the third (PLS-FS) and fourth (EN-FS) modeling approaches to develop BW prediction models.

The RMSEcv obtained for PLS-FS was around 53 kg (Table 2). The RPD ranged from 1.4 to 1.5. The relative predictive error ranged from 8.72% to 11.24% among the validation sets. The RMSEp estimated from those four validation sets was 60 ± 5.62 kg (Table 3). The trend of validation performance between datasets was similar to the one observed with PLS-NoMIR (Table 3). The absolute difference between RMSEcv and RMSEp was 7 kg (Table 3) which was lower than the one observed from models of PLS-NoMIR and PLS-All. Similar to PLS-All and PLS-NoMIR, the worse validation results were observed for herd10 (Table 2). The external validation on Australian data (herd12) had a RMSE of 64 ± 0.80 kg. The standard deviation is both lower than PLS-NoMIR and PLS-All. These same five selected predictors were also included in an EN regression. RMSEcv was similar for all four data sets and ranged between 52 and 53 kg (Table 2). The range of RPD and the relative error were from 1.4 to 1.5, and 8.72% to 11.50%, respectively. Validation performances were similar to PLS-FS but the SD of RMSEp and RMSEv were lower (Table 3). Therefore, these models were the optimal ones: low absolute difference between RMSEcv and RMSEp, low mean and SD of RMSEcv and RMSEPp and low mean and SD of RMSEv.

Besides the calibration and validation results, it is also interesting to assess the prediction consistency between the developed models. So, the 16 models were applied to the 4 validation datasets. The correlation coefficients among predictions using the 16 models applied on the four validation sets were high and ranged from 0.73 to 1.00. For all

Table 2
Calibration and validation performances obtained for the 16 developed models predicting test-day body weight (kg) of dairy cows.

Method	ValidationHerd	Herd5 (N = 149)	Herd7 (N = 188)	Herd10 (N = 180)	Herd11 (N = 135)
	Model number	1	2	3	4
PLS-NoMIR (Npredictors = 3;Parity, milk yield, and days in milk)	N samples	1699	1656	1662	1714
	N latent variables	2	2	2	2
	RMSEcv (kg)	53	53	52	53
	R^2cv	0.52	0.51	0.55	0.52
	RMSEp (kg)	60	55	68	58
	RMSEv (kg)	71	70	70	69
PLS-All (Npredictors = 280;;Parity, milk yield, days in milk, and 277 MIR spectral points)	N samples	1,699	1,658	1,663	1,716
	N latent variables	5	5	6	5
	RMSEcv (kg)	50	51	48	48
	R^2cv	0.59	0.55	0.63	0.62
	RMSEp (kg)	59	59	64	59
	RMSEv (kg)	64	63	93	126
PLS-FS (Npredictors = 5;Parity, milk yield, days in milk, and 2 MIR spectral points)	N samples	1696	1655	1661	1713
	N latent variables	2	2	2	2
	RMSEcv (kg)	53	53	52	53
	R^2cv	0.54	0.52	0.55	0.53
	RMSEp (kg)	61	54	67	57
	RMSEv (kg)	64	64	64	63
EN-FS (Npredictors = 5;Parity, milk yield, days in milk, and 2 MIR spectral points)	N samples	1696	1655	1661	1713
	Lambda	0.5	0.45	0.25	0.15
	RMSEcv (kg)	53	53	52	53
	R^2cv	0.54	0.52	0.55	0.53
	RMSEp (kg)	60	54	67	57
	RMSEv (kg)	65	64	64	63

PLS-NoMIR = Partial least squares regression (PLS) included milk yield, parity, and days in milk; PLS-All = PLS using milk yield, parity, days in milk and 277 MIR data; PLS-FS = PLS using the 5 variables selected by feature selection; EN-FS = Elastic Net regression using the 5 variables selected by feature selection; R^2cv = 10-fold cross-validation coefficient of determination; RMSEp = root mean square error of validation using herd5, herd7, herd10 or herd11; RMSEv = root mean square error estimated from herd12 (Australian data).

Table 3
Mean and standard deviation of cross-validation and validation root mean square errors (RMSE) per kind of models as well as the absolute difference between cross-validation RMSE (RMSEcv) and validation RMSE (RMSEp, RMSEv).

Models	RMSEcv		RMSEp		Diff_abs		RMSEv	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PLS-NoMIR	53	0.50	60	5.57	8	6.03	70	0.82
PLS-All	49	1.57	60	2.65	11	3.88	86	29.80
PLS-FS	53	0.50	60	5.62	7	6.05	64	0.80
EN-FS	53	0.50	60	5.57	7	6.02	64	0.69

RMSEcv = 10-fold cross-validation root mean square error; RMSEp = RMSE estimated from herd5, herd7, herd10, or herd11; Diff_abs denotes the absolute difference between RMSEcv and RMSEp; PLS-FS = PLS using the 5 variables selected by feature selection; EN-FS = Elastic Net regression using the 5 variables selected by feature selection.

calibration sets, the predictions of PLS-All models using the whole 280 predictors showed lower correlation coefficients (ranged from 0.73 to 0.96) with predictions obtained from PLS-NoMIR, PLS-FS and EN-FS models. Besides the lower RMSEp, RMSEv and their corresponding lower SD of models developed using EN-FS approach, the prediction correlation among EN-FS models ranged between 0.99 and 1 also revealing the robustness of this kind of modeling.

Table 4
Parameters of the variables selection algorithms based on PLS-SRD, PLS-UVE and Elastic Net.

Method[1]	ValidationHerd	Herd5 (N = 149)	Herd7 (N = 188)	Herd10 (N = 180)	Herd11 (N = 135)
	Number of Models	822	551	549	551
PLS-SRD	N samples	1,689	1,654	1,643	1,707
	N predictors	114	138	61	106
	N latent variables	5	5	6	5
PLS-UVE	N samples	1,696	1,657	1,661	1,712
	N predictors	106	120	69	116
	N latent variables	5	5	4	5
EN	N samples	1,696	1,656	1,663	1,711
	N predictors	67	62	73	62
	Penalty parameter	$\lambda = 0.5$	$\lambda = 0.5$	$\lambda = 0.5$	$\lambda = 0.5$

[1] PLS = Partial least squares regression; SRD = Sum of Ranking Difference; UVE = Uninformative Variable Elimination; EN = Elastic Net.

3.3. The robust model and variable importance

All EN-FS models presented a similar regression trend. Fig. 7A illustrates the relationships between predicted and real measured BW records (when herd7 as validation set). The slope of fitted regression line was 1.03, which was close to 1; the bias was −17.92 kg. For other calibration sets, the slope ranged between 0.98 and 1.00; the bias varied from 0.35 to 11.79 kg. No obvious herds presented a special distribution in the population cloud (Fig. 7A). The validation prediction errors were higher for the high and low BW (Fig. 7A). There was no obvious region revealed herds had extremely poor prediction at the BW range between 520 and 700 kg (Fig. 7A). To be more specific, the RMSEp and RMSEv were also estimated for 8 different 60 kg BW intervals (Fig. 7B). Clearly, the cows with BW lower than 400 kg or higher than 760 kg showed a relatively lower prediction accuracy. However, BW between 460 and 700 kg revealed a lower RMSE.

The variable relative importance estimated for the 5 selected variables included in the EN-FS model from the 4 datasets are given in Table 5. Based on the C value, the sequence of important variables was the parity, MIR0123, DIM, milk yield, and MIR0094. Those spectral points represent the MIR region of 1,396.21 cm^{-1} and 1,284.36 cm^{-1}. The variability of variable importance among models were low except for milk yield and MIR0094 (Table 5). Finally, according to the R^2 of the cross-validation step during the EN-FS regression based on the final 5 selected variable, the adding of MIR contributed on average 2.13% (R^2cv increased from around 0.53 to around 0.54) and up to 4.20% information to predict BW globally.

4. Discussion

4.1. The data sets structure and candidate predictors

To ensure the development of a robust model, the training dataset must be representative of the variability existing for the trait of interest (i.e., BW in this study) and the predictors should be sufficient informative. According to the mean and SD (Table 1), the average coefficient of variation for milk yield, fat, and protein contents were of 29.27%, 20.45% and 11.32%. This variability of the main milk components allowed indirectly to ensure a good milk spectral variability. The variability coefficient of BW varied between 6.30% and 14.36% in this study which was similar to the range of 8.17%–16.47% reported in literature (Koenen et al., 1999; Kuzuhara et al., 2015; McParland et al., 2015; Song et al., 2018; Yan et al., 2009, 2006). The datasets used can be therefore considered as representative of the cow population.

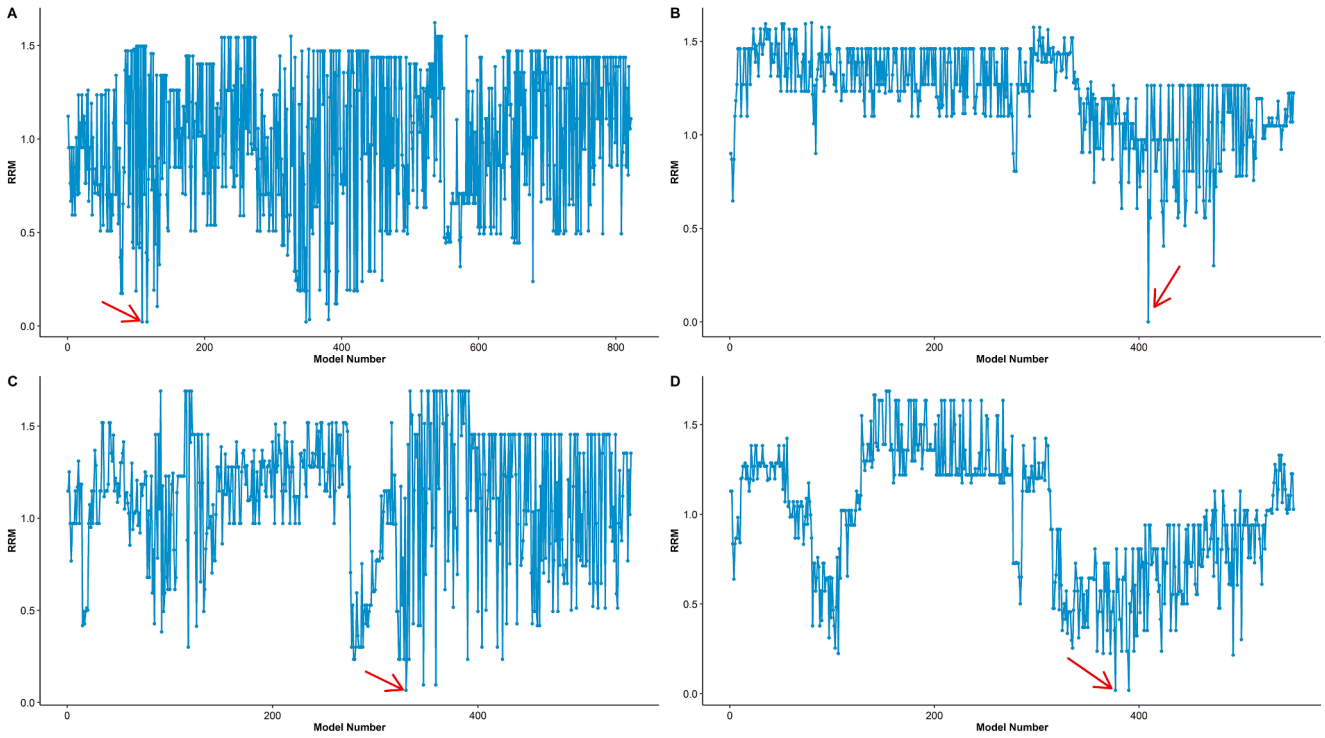

Fig. 5. The trend of RRM value across the model number for the 4 tested calibration sets (*A = first dataset until D = fourth dataset; the red arrow shows the optimal model*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

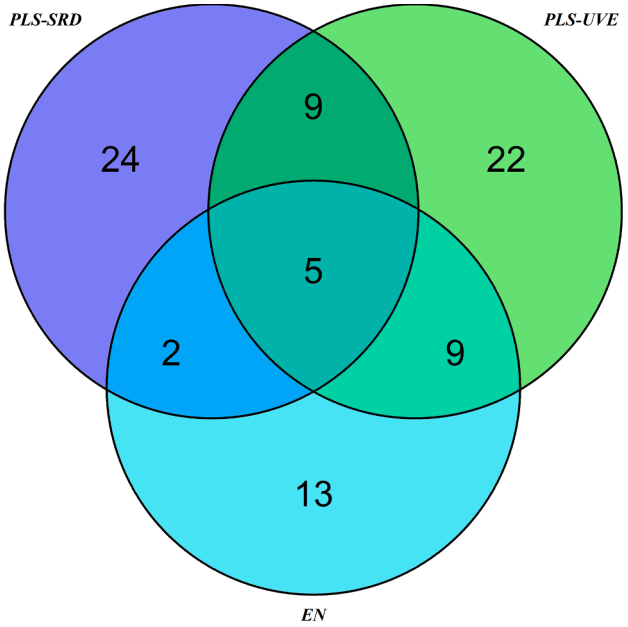

Fig. 6. Intersection of groups of selected variables among the PLS-SRD, PLS-UVE and EN feature selections and the four calibration sets.

Due to the interest of knowing the test-day BW at an individual level for management and breeding purposes, there is a need to try to predict this trait with predictors that are easily available on field by DHI organizations. So, it is why this study focused on the use of milk yield, DIM and parity as well as the milk MIR spectrum as predictors. These cow's characteristics related predictors were not chosen randomly. Indeed, these traits routinely recorded by DHI at cow level were positively related to BW. The correlation coefficients with measured BW were 0.12 for milk yield, 0.67 with parity, and 0.20 for DIM. Similar correlation values were already reported in the literature. For instance, Song et al. reported correlation coefficients of BW with parity and DIM equal to 0.81 and 0.18, respectively (Song et al., 2018). Although the correlation value between DIM and BW is not as high as the one observed with parity, this trait is valuable to be included in a prediction model as it is indirectly related to the pregnancy status of a cow (Song et al., 2018; Zhang et al., 2020). Moreover, the energy reallocation occurring after new calving cows (i.e., first days in milk) leads to a negative energy balance leading to a mobilization of body reserves (Soyeurt et al., 2019) influencing BW. Parity with its high correlation with BW is incontestably a relevant predictor. BW increased until the fourth parity and then stayed stable (Fig. 4D). This trend was expected as the cow's body keep growing until they got mature (i.e., normally stop at fourth or fifth parity) (Artegoitia et al., 2013; Mellado et al., 2011; Ray et al., 1992). Other authors reported the interest of using DIM and parity to predict BW as well (Enevoldsen and Kristensen, 1997; Kuzuhara et al., 2015). Along with the growth of cow's body, the increase of mammary volume from primiparous to multiparous of a healthy cow always result in an increase of milk yield (Davis and Hughson, 1988), explaining the positive relationship observed between milk yield and measured BW. The regression including those predictors (i.e., PLS-NoMIR) had a RMSEcv (53 kg; Table 3) similar to the value of 56 kg reported by Soyeurt et al. (2019) based on a model including parity, milk yield, DIM and test month. Even if these models seem to present satisfactory prediction performances on cross-validation and herd validation, the application of them on Australian data provided worse results suggesting a poor model transferability.

It is known that the milk composition can be related to the energy balance of the cow which is also related to the BW changes of a dairy cows (Friggens et al., 2007). Therefore, using the milk MIR spectral data reflecting the global milk composition as additional predictors is relevant to predict test-day BW. The modeling already proposed by Soyeurt et al. (2019) was not robust enough. Indeed, the absolute differences between RMSEp and RMSEcv were globally higher for PLS-All compared to the PLS-NoMIR models (Table 3). Moreover, the SD of RMSEcv and RMSEv were more than 3 times higher than the ones observed for PLS-

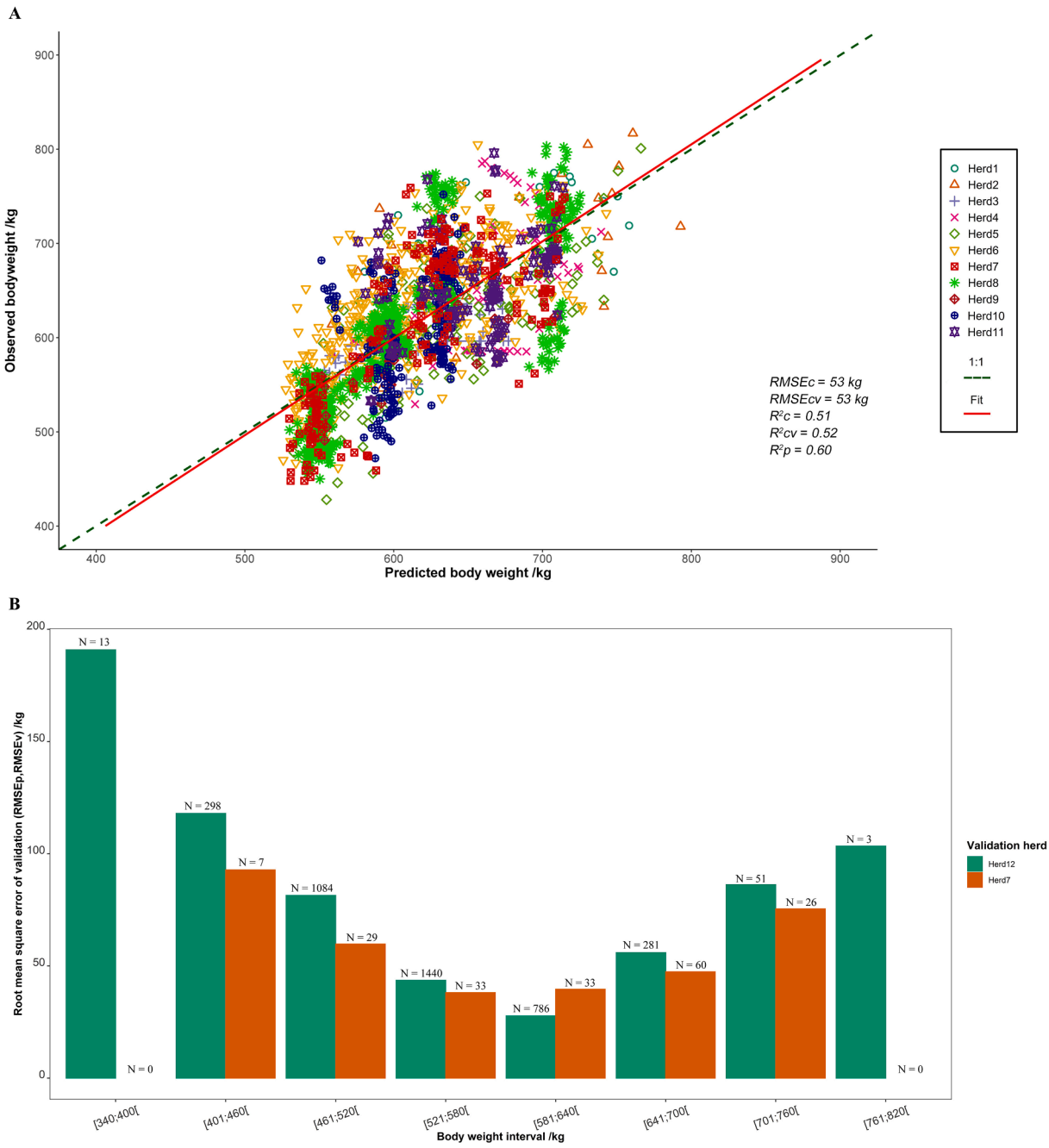

Fig. 7. The observed and predicted body weight of the calibration and validation herds (herd7) based on the Elastic Net regression using the final 5 selected variables (A); The validation root mean square error distribution across different body weight stage (interval = 60 kg; B).

Table 5
The relative importance of selected variables of the partial least squares regression models using different calibration sets.

Variables relative importance(%)	Herd5	Herd7	Herd10	Herd11	Mean	SD	CV	C value(Mean/SD)	Importance ranking
Parity	100.00	100.00	100.00	100.00	100.00	0.00	0.00	∞	1
MIR0123	12.11	12.49	11.15	11.67	11.86	0.57	4.84	2066.72	2
Days in milk (days)	40.88	41.38	43.74	45.74	42.93	2.25	5.24	1908.92	3
Milk yield (kg/d)	19.51	17.01	23.58	25.56	21.41	3.87	18.07	553.34	4
MIR0094	5.76	0.18	1.53	3.68	2.79	2.45	87.98	113.66	5

NoMIR suggesting the absence of robustness for those models. This could be related to the presence of uninformative variables in the model (Gottardo et al., 2016).

4.2. Feature selection and predictive model performance evaluation

4.2.1. Feature selection and the optimal predictors

Kalivas and Palmer (2014) reported that a harmonious model

requires a suitable balance of bias and variance to gain an optimal performance. However, using too few informative variables will lead to under-fit and the model is therefore not complex enough to capture the variability in the samples; whereas, adopting too many variables will result in an over-fitted model which is too specific to the calibration set used and presents therefore a poor predictive ability (Lavine, 2003). So, an optimum compromise must be found. In this context, a total of 3 different feature selection algorithms were employed in this study as the working hypothesis was that if a predictor is selected by different feature selection algorithms, the probability to highlight an informative variable is high. By using 4 different calibration sets, we have also reinforced this idea because the variable must be also selected from all datasets. By this way of doing, only 5 predictors were in common from the 12 selection procedures done in this study. There were then included in PLS and EN regressions. Both models (PLS-FS and EN-FS) had similar cross-validation and herd validation performances compared to PLS-NoMIR but a largely better validation performance on the Australian data (Table 3) suggesting a good transferability of those models. Moreover, as SD of RMSEv was slightly lower (while the SD of RMSEp was equal to PLS-NoMIR) for EN-FS, this model was assumed to be the most optimal model to predict test-day BW for dairy cows. In conclusion, the elimination of uninformative variables allows to keep the most informative variables and reduce the effect from noisy variables (Cai et al., 2008; Centner et al., 1996) leading to a better transferability of those models (i.e., mean and SD of RMSEv; Table 3). The low correlation found between predictions made using PLS-All approach and other approaches (i.e., using selected predictors) confirmed the presence of noisy information.

It is interesting to notice that the number of variables were different following the feature selection algorithms used and ranged from 61 to 138. However, at the end, by realizing the intersection of feature selection algorithms across training sets, only 5 variables (Parity, DIM, milk yield, and 2 MIR points) were used. This reduction of predictor number means that using only one feature selection algorithm and one dataset were not sufficient and would lead to an over-fitting of the model. Normally PLS algorithm should deal with that (Wold et al., 2001). Indeed, by fixing its LVs by taking into account the variability of predictors and the trait to be predicted, PLS could result in a low absolute regression coefficient of uninformative variable. However, this was not observed in this study. This could be related to the size of the dataset. Maybe by increasing the number of records, the PLS should improve its ability to recognize the most informative variables. This is why the use of a repeatability file (such as predicted BW as an indicator) composed of spectra without reference value could be interesting (Soyeurt et al., 2019; Vanrobays et al., 2015) and to test the selection of the whole MIR spectral points as well.

4.2.2. The predicted and observed body weights as well as the predictor importance

The relationship between the observed BW and the ones predicted using EN-FS records revealed the goodness of the calibration model (Fig. 7A). No obvious herd effect was visible on this graph. This can be explained by the high robustness of the EN-FS revealing by the small difference between RMSEp and RMSEcv. Furthermore, the information from the calibration herds have a good representativeness to the validation herd, especially in the models when leaving out herd5 and herd7 (lower RMSEp and RMSEv). Meanwhile, as we can observe from the fitted line of the predicted BW versus the real measured BW in Fig. 7A, the cows having extreme BW had a relative lower prediction performance. As reported by Thomas and Ge (2000), the representativeness and structure of a calibration set affect the accuracy of a predictive model. So, this phenomenon could be explained by the BW variability of the calibration set as there were a low amount of high and low BW in the calibration set (Fig. 4C). But the cows with a moderate BW had a better prediction accuracy.

The five selected variables were the parity, MIR0123, DIM, milk yield, and MIR0094. The mean and SD of relative variable importance values, and C value calculated from the EN-FS models showed that those first two predictors were the most important in all models. By averaging the relative variable importance values calculated for DIM, this predictor was the third more important and it was very close to MIR0123 which is related to the MIR region at 1,396 cm^{-1}. Not like the first three predictors, the variability of the milk yield relative variable importance values and the C value suggested that the importance of this trait can fluctuate between models. Similar trend was observed for the last predictor, MIR0094, which is related to the MIR region at 1284.36 cm^{-1}. These two MIR spectral data were located in the MIR spectral region related to the carbohydrates and organic acids (Bittante and Cecchinato, 2013; Picque et al., 1993). This could be related to the fact that the milk carbohydrate-related contents or fatty acids reflect partly the status of energy metabolism of body which influences the BW changes (Ducháček et al., 2012; Yan et al., 2006). The highlighting of parity, DIM and milk yield as important predictors of BW is in agreement with the findings mentioned in the first part of this discussion about the traits significantly related to BW (Enevoldsen and Kristensen, 1997; Kuzuhara et al., 2015; Song et al., 2018). The interest of using MIR information as additional predictors was confirmed in this study. In the past, McParland et al. (McParland et al., 2015) and Soyeurt et al. (2019) reported also the potential interest of using MIR spectra to predict BW changes or BW. However, in this study, the part of the explained BW variability brought by MIR data up to 4.20% was lower than the 11.67% reported by Soyeurt et al. (2019). Moreover, the current study revealed also the interest of adding MIR data to enhance the model robustness and transferability.

4.2.3. Accuracy of the predictive models

The RPD values for the EN-FS models averaged at 1.5, suggesting that the predicted BW is more an indicator of this trait. Indeed, a RPD between 1.5 and 2.0 representing a discrimination of low and high value of dependent variance (Saeys et al., 2005). However, although the value of RPD between 1.5 and 2.0 showed limited prediction accuracy, the prediction results could be accepted in a breeding view point (Chen et al., 2011). So, the RPD obtained in the current study around an average value of 1.5 indicating the usefulness of the models developed in this study to predict the body weight as least to distinguish the light and the heavier ones, and could be useful especially in the breeding programs. For instance, this predicted BW could be used in a breeding goal to measure the maintenance requirements (Koenen et al., 1999).

The RMSEp around 60 kg for the EN-FS representing 8.72%–11.46% of the real BW for different validation sets. This level of accuracy is acceptable, since, not like a certain research experiment, the measurement of BW in commercial farms, is also affected by many aspects like the measurement period (i.e., before or after milking of a cow, before or after feeding etc.) and the calibration of the weighing scale. In this study, no protocol was fixed to collect the BW records. Therefore, a certain uncertainty exists around the reference BW values used. However, it is important to notice that the accuracy obtained from 3-D camera or BCS regression methods ranged around 40 kg–50 kg regarding the cross-validation (Haile-Mariam et al., 2014; Kuzuhara et al., 2015). Although the prediction error isn't as small as the daily body weight changes (around 1 kg) resulted from body reserves mobilize (Jensen et al., 2015), whereas this extreme small changes appears normally during the early lactation stage (Gibb et al., 1992). So, in the current study, the accuracy of predicted BW is at least meaningful at global herd level. In this view, the mean live BW is often used in field to determine a global decision such as the nutritional requirement of a herd related to the goodness management of maintenance. So, from the breeding and herd level nutrient determination view, the residual of BW prediction, in the current study, using the selected models met the usage requirement practically.

5. Conclusion

This research confirmed the feasibility of predicting an indicator of test-day BW at an individual cow level using traits routinely recorded by DHI such as parity, days in milk, milk yield, and even animal conformation types. The use of the intersection of groups of variables using PLS-SRD, PLS-UVE and EN feature selection algorithms allowed to improve the model robustness and its transferability to another brand of spectrometers. The selection of parity, DIM, milk yield, MIR0094 and MIR0123 were logical based on the relationships between measured BW and those traits found in the literature. Even if PLS algorithm is normally able to deal with uninformative variables, this study showed the interest of realizing an elimination of uninformative variables before the final modeling. Moreover, in the context of this study, we demonstrated that the use of one feature selection algorithm and one data set was not sufficient to isolate the most informative variables. The EN-FS models developed in this study could be easily implemented into a routine DHI framework allowing to the creation of a large level database of BW indicator. However, the herd specific BW evolution features could be considered in the future studies.

CRediT authorship contribution statement

Lei Zhang: Conceptualization, Methodology, Software, Investigation, Visualization, Writing - original draft, Writing - review & editing, Formal analysis, Project administration, Data curation, Resources. **Anthony Tedde:** Software, Writing - review & editing. **Phuong Ho:** Data curation, Resources, Writing - review & editing. **Clément Grelet:** Project administration, Data curation, Resources, Formal analysis, Writing - review & editing. **Frédéric Dehareng:** Project administration, Data curation, Resources, Writing - review & editing. **Eric Froidmont:** Data curation, Resources, Writing - review & editing. **Nicolas Gengler:** Methodology, Project administration, Data curation, Resources, Writing - review & editing. **Yves Brostaux:** Formal analysis, Writing - review & editing. **Dagnachew Hailemariam:** Data curation, Resources, Writing - review & editing. **Jennie Pryce:** Data curation, Resources, Writing - review & editing. **Hélène Soyeurt:** Methodology, Formal analysis, Project administration, Data curation, Resources, Conceptualization, Software, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The China Scholarship Council (**CSC**) is gratefully acknowledged for funding Mr. Zhang's PhD grant.

The members of GplusE Consortium which including: Mark Crowe, Alan Fahey, Fiona Carter, Elizabeth Matthews, Andreia Santoro, Colin Byrne, Pauline Rudd, Roisin O'Flaherty, Sinead Hallinan, Claire Wathes, Mazdak Salavati, Zhangrui Cheng, Ali Fouladi, Geoff Pollott, Dirk Werling, Beatriz Sanz Bernardo, Conrad Ferris, Alistair Wylie, Matt Bell, Mieke Vaneetvelde, Kristof Hermans, Miel Hostens, Geert Opsomer, Sander Moerman, Jenne De Koster, Hannes Bogaert, Jan Vandepitte, Leila Vandevelde, Bonny Vanranst, Klaus Ingvartsen, Martin Tang Sorensen, Johanna Hoglund, Susanne Dahl, Soren Ostergaard, Janne Rothmann, Mogens Krogh, Else Meyer, Leslie Foldager, Charlotte Gaillard, Jehan Ettema, Tine Rousing, Torben Larsen, Victor H. Silva de Oliveira, Cinzia Marchitelli, Federica Signorelli, Francesco Napolitano, Bianca Moioli, Alessandra Crisà, Luca Buttazzoni, Jennifer McClure, Daragh Matthews, Francis Kearney, Andrew Cromie, Matt McClure, Shujun Zhang, Xing Chen, Huanchun Chen, Junlong Zhao, Liguo Yang, Guohua Hua, Chen Tan, Guiqiang Wang, Michel Bonneau, Marlène Sciarretta, Armin Pearn, Arnold Evertson, Linda Kosten, Anders Fogh, Thomas Andersen, Matthew Lucy, Chris Elsik, Gavin Conant, Jerry Taylor, Deborah Triant, Nicolas Gengler, Michel Georges, Frederic Colinet, Marilou Ramos Pamplona, Hedi Hammami, Catherine Bastin, Haruko Takeda, Aurelie Laine, Anne-Sophie Van Laere, Rodrigo Mota, Saied Naderi Darbagshahi, Frederic Dehareng, Clement Grelet, Amelie Vanlierde, Eric Froidmont, Frank Becker, Martin Schulze, and Sergio Palma Vera are deeply acknowledged in this study.

The authors gratefully thank the farmers and staff of milk recording organizations for the collection of data used in this study: Ellinbank (Australia), University of liege (Liege, Belgium), Walloon Agricultural Research Centre (Gembloux, Belgium), University of Alberta (Alberta, Canada), Aarhus University (Tjele, Denmark), Agri-Food and Biosciences Institute (Northern Ireland), University College Dublin (Dublin, Ireland), Walloon Breeding Association (Ciney, Belgium), and Leibniz Institute for Farm Animal Biology (Dummerstorf, Germany).

Funding

The Genotype Plus Environment (GplusE) Project has received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 613689. The authors also acknowledge the National Fund for Scientific Research for its support through the PDR SimBa project (T022119F).

Ethics declarations

This work was carried out in accordance with the EU Directive 2010/63/EU for animal experiments.

References

Artegoitia, V., Meikle, A., Olazabal, L., Damián, J.P., Adrien, M.L., Mattiauda, D.A., Bermudez, J., Torre, A., Carriquiry, M., 2013. Milk casein and fatty acid fractions in early lactation are affected by nutritional regulation of body condition score at the beginning of the transition period in primiparous and multiparous cows under grazing conditions. J. Anim. Physiol. Anim. Nutr. (Berl.) 97, 919–932.

Banos, G., Coffey, M.P., 2012. Technical note: Prediction of liveweight from linear conformation traits in dairy cattle. J. Dairy Sci. 95, 2170–2175. https://doi.org/10.3168/jds.2011-4838.

Bewley, J.M., Schutz, M.M., 2008. An Interdisciplinary Review of Body Condition Scoring for Dairy Cattle. Prof. Anim. Sci. 24, 507–529. https://doi.org/10.15232/s1080-7446(15)30901-3.

Bittante, G., Cecchinato, A., 2013. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds. J. Dairy Sci. 96, 5991–6006. https://doi.org/10.3168/jds.2013-6583.

Bjørn-Helge Mevik, R.W., K.H.L., 2019. Partial Least Squares and Principal Component Regression. R package version 2.7-2.

Cai, W., Li, Y., Shao, X., 2008. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemom. Intell. Lab. Syst. 90, 188–194. https://doi.org/10.1016/j.chemolab.2007.10.001.

Centner, V., Massart, D.-L., de Noord, O.E., de Jong, S., Vandeginste, B.M., Sterna, C., 1996. Elimination of Uninformative Variables for Multivariate Calibration. Anal. Chem. 68, 3851–3858. https://doi.org/10.1021/ac960321m.

Chen, G.L., Zhang, B., Wu, J.G., Shi, C.H., 2011. Nondestructive assessment of amino acid composition in rapeseed meal based on intact seeds by near-infrared reflectance spectroscopy. Anim. Feed Sci. Technol. 165, 111–119. https://doi.org/10.1016/j.anifeedsci.2011.02.004.

Davies, A.M.C., Fearn, T., 2006. Back to basics: calibration statistics. Spectrosc. Eur. 18, 31–32.

Davis, S.R., Hughson, G.A., 1988. Measurement of functional udder capacity in lactating Jersey cows. Aust. J. Agric. Res. 39, 1163–1168.

Edmonson, A.J., Lean, I.J., Weaver, L.D., Farver, T., Webster, G., 1989. A Body Condition Scoring Chart for Holstein Dairy Cows. J. Dairy Sci. 72, 68–78. https://doi.org/10.3168/jds.S0022-0302(89)79081-0.

Enevoldsen, C., Kristensen, T., 1997. Estimation of body weight from body size measurements and body condition scores in dairy cows. J. Dairy Sci. 80, 1988–1995.

Friggens, N.C., Ridder, C., Løvendahl, P., 2007. On the use of milk composition measures to predict the energy balance of dairy cows. J. Dairy Sci. 90, 5453–5467. https://doi.org/10.3168/jds.2006-821.

Gibb, M.J., Ivings, W.E., Dhanoa, M.S., Sutton, J.D., 1992. Changes in body components of autumn-calving Holstein-Friesian cows over the first 29 weeks of lactation. Anim. Sci. 55, 339–360.

Gottardo, P., Penasa, M., Lopez-Villalobos, N., De Marchi, M., 2016. Variable selection procedures before partial least squares regression enhance the accuracy of milk fatty acid composition predicted by mid-infrared spectroscopy. J. Dairy Sci. 99, 7782–7790.

Gowen, A.A., Downey, G., Esquerre, C., O'Donnell, C.P., 2011. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. J. Chemom. 25, 375–381. https://doi.org/10.1002/cem.1349.

Grelet, C., Froidmont, E., Foldager, L., Salavati, M., Hostens, M., Ferris, C.P., Ingvartsen, K.L., Crowe, M.A., Sorensen, M.T., Fernandez Pierna, J.A., Vanlierde, A., Gengler, N., Dehareng, F., 2020. Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation. J. Dairy Sci. https://doi.org/10.3168/jds.2019-17910.

Grelet, C., Pierna, J.A.F., Dardenne, P., Soyeurt, H., Vanlierde, A., Colinet, F., Bastin, C., Gengler, N., Baeten, V., Dehareng, F., 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 100, 7910–7921.

Haile-Mariam, M., Gonzalez-Recio, O., Pryce, J.E., 2014. Prediction of liveweight of cows from type traits and its relationship with production and fitness traits. J. Dairy Sci. 97, 3173–3189. https://doi.org/10.3168/jds.2013-7516.

Ho, P.N., Marett, L.C., Wales, W.J., Axford, M., Oakes, E.M., Pryce, J.E., 2019. Predicting milk fatty acids and energy balance of dairy cows in Australia using milk mid-infrared spectroscopy. Anim. Prod. Sci. 60, 164–168. https://doi.org/10.1071/AN18532.

ICAR, 2017. Section 2 - Guidelines for Dairy Cattle Milk Recording, in: Guidelines. p. 27.

Jensen, C., Østergaard, S., Bertilsson, J., Weisbjerg, M.R., 2015. Responses in live weight change to net energy intake in dairy cows. Livest. Sci. 181, 163–170. https://doi.org/10.1016/j.livsci.2015.09.016.

Kalivas, J.H., Héberger, K., Andries, E., 2015. Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. Anal. Chim. Acta 869, 21–33.

Kalivas, J.H., Palmer, J., 2014. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. J. Chemom. 28, 347–357. https://doi.org/10.1002/cem.2555.

Koenen, E.P.C., Groen, A.F., Gengler, N., 1999. Phenotypic variation in live weight and live-weight changes of lactating Holstein-Friesian cows. Anim. Sci. 68, 109–114. https://doi.org/10.1017/s135772980005013x.

Korver, S., 1988. Genetic aspects of feed intake and feed efficiency in dairy cattle: a review. Livest. Prod. Sci. 20, 1–13.

Kuzuhara, Y., Kawamura, K., Yoshitoshi, R., Tamaki, T., Sugai, S., Ikegami, M., Kurokawa, Y., Obitsu, T., Okita, M., Sugino, T., 2015. A preliminarily study for predicting body weight and milk properties in lactating Holstein cows using a three-dimensional camera system. Comput. Electron. Agric. 111, 186–193.

Lavine, B., 2003. A user-friendly guide to multivariate calibration and classification, Tomas Naes, Tomas Isakson, Tom Fearn and Tony Davies, NIR Publications, Chichester, 2002. J. Chemom. 17, 571–572. https://doi.org/10.1002/cem.815.

Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta 648, 77–84.

Maltz, E., Devir, S., Metz, J.H.M., Hogeveen, H., 1997. The body weight of the dairy cow: I. Introductory study into body weight changes in dairy cows as a management aid. Livest. Prod. Sci. 48, 175–186. https://doi.org/10.1016/S0301-6226(97)00024-9.

Mäntysaari, P., Mäntysaari, E.A., 2015. Modeling of daily body weights and body weight changes of Nordic Red cows. J. Dairy Sci. 98, 6992–7002. https://doi.org/10.3168/jds.2015-9541.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, C.C. and T.H., 2019. caret: Classification and Regression Training.

McParland, S., Kennedy, E., Lewis, E., Moore, S.G., McCarthy, B., O'Donovan, M., Berry, D.P., 2015. Genetic parameters of dairy cow energy intake and body energy status predicted using mid-infrared spectrometry of milk. J. Dairy Sci. 98, 1310–1320. https://doi.org/10.3168/jds.2014-8892.

Mehmood, T., Liland, K.H., Snipen, L., Sæbø, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemom. Intell. Lab. Syst. 118, 62–69.

Mehmood, T., Sæbø, S., Liland, K.H., 2020. Comparison of variable selection methods in partial least squares regression. J. Chemom. n/a, e3226 https://doi.org/10.1002/cem.3226.

Mellado, M., Antonio-Chirino, E., Meza-Herrera, C., Veliz, F.G., Arevalo, J.R., Mellado, J., De Santiago, A., 2011. Effect of lactation number, year, and season of initiation of lactation on milk yield of cows hormonally induced into lactation and treated with recombinant bovine somatotropin. J. Dairy Sci. 94, 4524–4530.

Miglior, F., Fleming, A., Malchiodi, F., Brito, L.F., Martin, P., Baes, C.F., 2017. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. J. Dairy Sci. 100, 10251–10271.

Müllertz, A., Perrie, Y., Rades, T., 2016. Analytical techniques in the pharmaceutical sciences. Springer.

Nie, M., Meng, L., Chen, X., Hu, X., Li, L., Yuan, L., Shi, W., 2019. Tuning parameter identification for variable selection algorithm using the sum of ranking differences algorithm. J. Chemom. 33, e3113 https://doi.org/10.1002/cem.3113.

Picque, D., Lefier, D., Grappin, R., Corrieu, G., 1993. Monitoring of fermentation by infrared spectrometry: Alcoholic and lactic fermentations. Anal. Chim. Acta 279, 67–72. https://doi.org/10.1016/0003-2670(93)85067-T.

R Core Team, 2019. A language and environment for statistical computing. R Foundation for Statistical Computing.

Ray, D.E., Halbach, T.J., Armstrong, D.V., 1992. Season and lactation number effects on milk production and reproduction of dairy cattle in Arizona. J. Dairy Sci. 75, 2976–2983. https://doi.org/10.3168/jds.S0022-0302(92)78061-8.

Saeys, W., Mouazen, A.M., Ramon, H., 2005. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. Biosyst. Eng. 91, 393–402. https://doi.org/10.1016/j.biosystemseng.2005.05.001.

Song, X., Bokkers, E.A.M., van der Tol, P.P.J., Koerkamp, P.W.G.G., van Mourik, S., 2018. Automated body weight prediction of dairy cows using 3-dimensional vision. J. Dairy Sci. 101, 4448–4459.

Soyeurt, Hélène, Froidmont, E., Dufrasne, I., Hailemariam, D., Wang, Z., Bertozzi, C., Colinet, F.G., Dehareng, F., Gengler, N., 2019. Contribution of milk mid-infrared spectrum to improve the accuracy of test-day body weight predicted from stage, lactation number, month of test and milk yield. Livest. Sci. 227, 82–89.

Tencate, A.J., Kalivas, J.H., White, A.J., 2016. Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: A model updating application for pharmaceutical analysis. Anal. Chim. Acta 921, 28–37. https://doi.org/10.1016/j.aca.2016.03.046.

Thomas, E.V., Ge, N., 2000. Development of Robust Multivariate Calibration Models. Technometrics 42, 168–177. https://doi.org/10.1080/00401706.2000.10485996.

Thorup, V.M., Edwards, D., Friggens, N.C., 2012. On-farm estimation of energy balance in dairy cows using only frequent body weight measurements and body condition score. J. Dairy Sci. 95, 1784–1793. https://doi.org/10.3168/jds.2011-4631.

Ducháček, J., Vacek, M., Stádník, L., Beran, J., Okrouhlá, M., 2012. Changes in milk fatty acid composition in relation to indicators of energy balance in Holstein cows. Acta Univ. Agric. Silvic. Mendelianae Brun. 60, 29–38. https://doi.org/10.11118/actaun201260010029.

Vanrobays, M.L., Vandenplas, J., Hammami, H., Froidmont, E., Gengler, N., 2015. Short communication: Novel method to predict body weight of primiparous dairy cows throughout the lactation. J. Dairy Sci. 98, 692–697. https://doi.org/10.3168/jds.2014-8504.

Wang, Y., Jiang, F., Gupta, B.B., Rho, S., Liu, Q., Hou, H., Jing, D., Shen, W., 2018. Variable Selection and Optimization in Rapid Detection of Soybean Straw Biomass Based on CARS. IEEE Access 6, 5290–5299. https://doi.org/10.1109/ACCESS.2017.2763596.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Yan, T., Mayne, C.S., Keady, T.W.J., Agnew, R.E., 2006. Effects of Dairy Cow Genotype with Two Planes of Nutrition on Energy Partitioning Between Milk and Body Tissue. J. Dairy Sci. 89, 1031–1042. https://doi.org/10.3168/jds.S0022-0302(06)72170-1.

Yan, T., Mayne, C.S., Patterson, D.C., Agnew, R.E., 2009. Prediction of body weight and empty body composition using body size measurements in lactating dairy cows. Livest. Sci. 124, 233–241. https://doi.org/10.1016/j.livsci.2009.02.003.

Zhang, L., Gengler, N., Dehareng, F., Colinet, F., Froidmont, E., Soyeurt, H., 2020. Can we observe expected behaviors at large and individual scales for feed efficiency-related traits predicted partly from milk mid-infrared spectra? Animals 10, 873. https://doi.org/10.3390/ani10050873.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol) 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.