Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Selection of reference samples for updating multivariate calibration models used in the analysis of pig faeces

Andrés Cruz-Conesa [a,b], Joan Ferré [a,*], Itziar Ruisánchez [a], Anna M. Pérez-Vendrell [b]

[a] Universitat Rovira i Virgili, Faculty of Chemistry, Department of Analytical and Organic Chemistry, Tarragona, Spain
[b] Institute of Agrifood Research and Technology, Animal Nutrition, Mas Bové, Constantí, Spain

ARTICLE INFO

Keywords:
Sample selection
Model updating
Near-infrared spectroscopy
D-optimal
Phosphorus
PLS

ABSTRACT

Monitoring and updating calibration models are common tasks when analytical methods are based on near-infrared spectroscopy. This work describes a situation in which a PLS calibration model that is used routinely for the determination of phosphorus content in pig faeces in digestibility studies had to be updated in order to be used with the faeces collected in a new trial with phytases. An approach based on D-optimality is presented that selects a reduced number of the new samples to be analyzed with the reference analytical method so that the small set is used to confirm the need to update the model and validate it. The rest of the new samples that had not been selected by the algorithm were accurately predicted with the updated model. The updated model maintained its previous performance for the samples in the validation set (an RMSEP of 1.58 g kg^{-1} compared with an RMSEP of 1.54 g kg^{-1} before the update) and the prediction error for the new samples was RMSECV = 1.95 g kg^{-1}, much lower than the RMSEP = 11.38 g kg^{-1} obtained before the model update. In addition, the predictive ability of the updated PLS model was significantly better than updated models selecting the reduced dataset using other sample selection methods such as Kennard-Stone, a leverage-based selection method and random selection.

1. Introduction

Animal nutrition research is a wide field aimed at the efficient and sustainable production of food. A large branch of animal nutrition research is devoted to finding optimal formulations for the diets of farm animals at the different growth stages and understanding how the ingredients interact and enhance nutrient digestibility. This valuable information is obtained in *in-vivo* trials in which ingredients, feeds and faeces must be analyzed. Over the years, the traditional time-consuming analytical methods used to analyze these samples have been replaced by rapid, reagent-free, waste-free determinations based on near-infrared spectroscopy (NIRS) and multivariate calibration. These models have been shown to predict accurately the nutritional content of a variety of ingredients [1–4], feeds [5,6] and faeces [7,8] so NIRS is now widely used, not only in animal research but also as a routine analytical technique and legal feed labelling by feed producers in the agri-food sector.

To provide accurate predictions, NIRS-based models must be trained on representative samples for which the reference parameters have been determined with validated analytical methods, usually official methods. This is the longest part of method development since it involves collecting and analysing samples from different sources, over long periods of time until all probable future sources of spectral variations have been taken into account. In the feed-production sector, for example, this implies including different raw materials from different origins, from various harvests and stored in a variety of conditions. Since model predictions are only reliable for samples obtained under the same conditions (within the limits) as the samples from the calibration set, the performance of these models must be monitored to uncover unmodelled spectral variability. If it exists, then the model must be adapted to the new situation. To do that, some calibration transfer or domain adaptation methods have been proposed in the literature [9,10]. Among them, the present study focuses on model update with new samples to expand the domain [11]. How often an update is needed depends on how universal the model is. It is less frequent in feed production facilities, which use stable sources of raw materials, than in animal research studies, where new combinations of ingredients are regularly being tested. This work describes an example of the latter, in which NIRS-based multivariate models are used to predict organic matter, crude protein, fat, neutral detergent fibre, acid detergent fibre and phosphorus in pig faeces during digestibility studies carried out at the Institute of Agrifood

[*] Corresponding author.
E-mail address: joan.ferre@urv.cat (J. Ferré).

https://doi.org/10.1016/j.chemolab.2022.104749
Received 21 October 2022; Received in revised form 17 December 2022; Accepted 29 December 2022
Available online 30 December 2022

Research and Technology (IRTA) in Constantí, Tarragona, Spain. The composition of the studied faeces depends on the weight, age, and genetic background of the animal [12] but especially on the type and digestibility of the diet. Therefore, it is not uncommon for faeces spectra from a new digestibility trial to show variations not recognized by the current model. If that were the case, the model would produce unreliable predictions for these samples and therefore they would have to be analyzed with the slower analytical method. Alternatively, one wishes not to analyze the whole batch of samples but just a reduced representative number of them. This subset would be then used to update the calibration model so that the new model can be used to predict the rest of the batch as well as all future samples of the same type.

The selection of representative samples is a recurrent topic in the NIRS literature, ranging from dividing a dataset into training, validation and test sets [13], to selecting a subset of samples for model transfer between instruments or in new conditions [14], as well as the selection of samples for updating running models [11]. This work focuses on the latter case. The simplest selection method in model updating is to randomly select samples from the new batch. While this method is statistically sound, the main drawback is that the selected subset may not expand to the limits of the new spectral domain. Hence, some of the samples to be predicted may still appear extreme to the updated model, leading to extrapolation problems [15]. To better ensure the representativeness of the selected subset and the coverage of the spectral domain, the selection can be made with specific algorithms. Algorithms can be grouped into those that make use of the spectra (X) and the reference values (y) and those that only use the spectra. Sample set partitioning based on the joint X-y distances (SPXY) is an example of the first group [16]. This algorithm has been shown to select more representative subsets than those based solely on the spectra [17]. Other examples are the successive projections algorithm (SPA) [18] and the reference value (YR)-based sample selection algorithm [19]. The main drawback of these algorithms is that they require analysing all the new samples with the reference method. This means that the updated model will only be useful for future samples but the batch that just arrived cannot benefit from the update. A more interesting approach is to select samples based on the spectra only, which is particularly useful when the new batch is large. Once the outlier diagnostics have warned about the likely inaccuracy of the predictions, a few of these samples are selected based on their spectra, are analyzed by the reference method and are used to update the model. The updated model can be then used to predict the rest of the batch. A variety of algorithms can select the samples based solely on the spectra. Two popular ones are the Kennard-Stone [20] and the duplex [21] algorithms that try to uniformly cover the multidimensional spectral space by selecting the samples with the maximum distance (commonly Euclidean distance or Mahalanobis distance) between the selected samples. Other algorithms are those based on leverage or Mahalanobis distance [22,23]. More recently, Xu et al. [24] proposed to use the simple interval calculation (SIC) leverage as the criterion and Chen et al. [25] developed an algorithm based on isolation forests for outlier detection and subset selection (IOS). The cited algorithms rank the samples by their importance but there is no criterion that indicates what is a sufficient number of samples. An optimal subset size can be provided by criteria from the field of optimal design of experiments. To decide optimal sets of experimental conditions, criteria such as the D-criterion, the G-efficiency criterion and the A-criterion [26] provide optimal sets on the basis of Multiple Linear Regression (MLR) models. Ferré and Rius used the D-optimality criterion to select samples based on their spectra for different types of spectroscopies [27,28]. In their work, the samples were selected from an initial set for building a model, but not to update an existing one.

This work presents as a case study the selection of samples for the updating of a partial least squares (PLS) model for the prediction of phosphorus content in pig faeces from their NIR spectra. The need arose when the spectra of the pig faeces from a new digestibility trial were flagged as outliers for the running model. Thus, a reduced subset of samples from the new trial was selected, analyzed with the reference method and used to update the model. The sample selection approach is inspired by the D-optimality criterion used in optimal experimental design. Its performance is compared with that of random selection, the Kennard-Stone algorithm and selection based on the leverage.

2. Materials and methods

2.1. Samples

The existing model for phosphorus content in pig faeces was built with pig faeces samples collected during digestibility studies from 2018 to 2020 at the Institute of Agrifood Research and Technology (IRTA) in Constantí, Spain. The faeces samples were lyophilized, ground and stored in sealed bags in the refrigerator until analysis. Phosphorus content (expressed as g kg^{-1} related to raw product) was determined by UV-VIS spectroscopy using molybdovanadate reagent according to AOAC Official Method 965.17 [29]. The spectrum of approximately 30 g of sample was measured on a NIRS DS2500 (Foss NIR Systems, Denmark) with a 7 cm diameter cup in reflectance mode from 800 to 2499.5 nm every 0.5 nm. The data set was randomly divided into 246 training samples and 83 validation samples that were used to develop the model running in the laboratory. In 2021, a new batch of 103 samples was collected in a digestibility study that investigated the efficacy of different phytases on the performance of weaned piglets fed with a complex diet based on wheat-corn and soybean meal. The determination of phosphorus in these samples from their spectra using the existing calibration model produced outlier detection warnings and hence, an update of the current model was required. For this process, the phosphorus content in some selected samples of the new batch was needed. This was found with the reference method for phosphorus content mentioned above.

2.2. Data analysis

Partial least squares regression (PLSR) was used to develop the calibration model for phosphorus content. The spectra were pretreated with Savitzky–Golay first derivative interpolating with a second-order polynomial and a window width of 17 points [30] and mean-centered. A five-fold venetian blind cross validation was used to choose the optimal number of latent variables (LVs) for the model. PLS toolbox software (2016, Eigenvector Research, Inc., Manson, WA, USA) running in Matlab R2020a (The MathWorks Inc., Natick, MA, USA) was used to develop the calibration models. The Fedorov algorithm used for sample selection was programmed in-house.

2.3. Selection of samples for model updating

D-optimality is a numerical criterion used in the field of optimal experimental design that defines the quality of an experimental design. Standard experimental designs, such as the full factorial design, are D-optimal. D-optimality is also used to create experimental designs when the experimental domain is irregular, in which case an algorithm such as Fedorov's algorithm [31] or genetic algorithms [32] is used to choose from a list of candidate examples which ones should be selected, under the criterion that they should maximize the determinant of the information matrix in an MLR model. In the context of multivariate calibration, this idea was used to select subsets of calibration samples to obtain models using fewer samples. It was found to perform as well as or even better than Kennard-Stone algorithm or random sampling [27,33,34]. Different from the Kennard–Stone algorithm, which seeks to maximize the distance between the selected samples to uniformly cover the calibration domain, the samples selected with the D-optimal criterion tend to be located at the edges of the domain in lineal models where the most influential samples are found.

D-optimality is also used in experimental design to repair designs when some of the planned experiments cannot be executed, or the domain must be extended. In that case, the algorithm searches which examples from a list of candidates should be added to the existing ones in order to maximize the determinant of the information matrix. This use of D-optimality is exploited in this work, where a procedure using Fedorov's algorithm is presented. This algorithm will be used to find which samples from an external trial should be added to the existing calibration set to update the original PLS model.

For a working PLS model, let $\mathbf{T}$ be the $N \times P$ matrix of scores of the calibration samples where each row corresponds to a calibration sample and each column corresponds to a latent variable of the model. Let $\mathbf{T}_\mathrm{B}$ be $I \times P$ matrix of the scores of the new batch of samples projected onto the latent variable space of the current model. These samples are candidates to being analyzed and used for model updating. The algorithm starts by creating matrix $\mathbf{T}_n$ by randomly selecting n rows of $\mathbf{T}_\mathrm{B}$. Then, the determinant $\mathrm{Det}(\mathbf{T}_\mathrm{E}^\mathrm{T}\,\mathbf{T}_\mathrm{E})$ is evaluated, where $\mathbf{T}_\mathrm{E}$ is the matrix $\begin{bmatrix}\mathbf{T}\\ \mathbf{T}_n\end{bmatrix}$ and T indicates transpose. Next, one of the rows of $\mathbf{T}_n$ is exchanged for one of the remaining rows of $\mathbf{T}_\mathrm{B}$ so that the determinant increases as much as possible. The exchanged rows are decided following Fedorov's algorithm to find D-optimal subsets and can be found elsewhere [26–28]. This step is repeated iteratively until the determinant no longer improves. Since the algorithm can find local maxima, it can be restarted multiple times with a new random set of samples $\mathbf{T}_n$ and the set of n samples with the largest determinant is kept as optimal. The whole procedure is then repeated to select subsets with a different number of samples n and the one with a maximum $\mathrm{Det}\left(\frac{\mathbf{T}_\mathrm{E}^\mathrm{T}\,\mathbf{T}_\mathrm{E}}{n+N}\right)^{\frac{1}{p}}$, which is a measure of the information content per sample, is finally selected as the optimal subset to be used for updating the model.

3. Results and discussion

3.1. Detecting the need for model updating

At IRTA, a PLSR model was used to predict the phosphorus content in faeces. Table 1 shows the performance measures for this model. The model involved 14 LVs as determined by cross-validation. This relatively high number was attributed to the fact that the training set included faeces from diverse research studies from 2018 to 2020, which involved different diets fed to pigs of different ages, sexes and weights. Also notice that different from other nutrients such as protein or fat, phosphorus does not present specific absorption bands in the studied spectral range. Its prediction is possible thanks to the correlations between this constituent and the absorbance of some organic molecules of the sample such as phytate that it is abundant in plant materials and hence in pig diets and faeces. This makes the prediction of phosphorus usually less accurate than the predictions for main nutrients that have stronger contributions in NIRS. Overall, the model performed well with a coefficient of determination of prediction (R_p^2) of 0.94 and a root mean square error of prediction (RMSEP) of 1.54 g kg^{-1} for the validation set. Thus, the model was considered valid for routine use. The routine use of the model included checking the sum of squares of the spectral residuals (Q-residuals) and the leverage of the new samples to be predicted [35]. These are common diagnostics to flag unmodeled spectral variability in the new spectra. Q-residuals and/or leverage larger than those of the training and validation samples warn of unreliable predictions. The detected outliers can be the result of erroneous measurements but can also indicate unique samples. The analyst's decision on how to proceed next depends on knowledge about the unusual characteristics of the new samples and whether they are occasional (so they should simply be analyzed with the reference method) or whether more samples of the same type are likely to arrive in the future, in which case it may be worth updating the model [36].

In this work, a new batch of pig faeces came from a trial coded P704 that tested a complex diet with ingredients that had not been used before. This included rapeseed meal, rice bran and sunflower seeds, but also different phytases, which could affect the digestion of the nutrients in the feed, mainly phosphorus digestion. Therefore, it was likely that the faeces from this trial could be outliers for the current model for phosphorus content. As expected, the faeces spectra had significant differences in signal intensity from those used to train and validate the PLS model (Fig. 1). In addition, the entire batch was positioned in the upper right quadrant of the Q-residuals versus leverage plot (Fig. 2). The unmodeled variability in the new spectra could lead to large errors in the predicted phosphorus content. Note that although limits for the leverage and Q-residuals can be defined from the training and validation data [35,37], in this case visual inspection of this plot was sufficient to reveal the abnormal behavior of the new batch. The expected large errors were confirmed by analyzing a subset of the samples of trial P704 with the reference method. As expected, (Fig. 3), the errors were unacceptably large, with an RMSEP of 11.38 g kg^{-1}, much higher than the 1.54 g kg^{-1}

Table 1
Characteristics of the previous and updated calibration models developed for phosphorus content in pig faeces (g·kg^{-1}). Number of samples used for calibration (N_c) and validation (N_v), number of P704 trial samples in the D-optimal selected subset ($N_{\mathrm{D-sel}}$), number of samples used to validate the methodology ($N_{D-\mathrm{val}}$), number of latent variables (LV), coefficient of determination of calibration (R_c^2) and prediction (R_p^2), root mean square error of calibration (RMSEC) and prediction (RMSEP), bias and slope of the predicted vs measured regression line, root mean square error of prediction of the P704 samples included in the subset (*RMSEP*$_{\mathrm{D-sel}}$) and used to validate the methodology (*RMSEP*$_{\mathrm{D-val}}$).

	Previous model	Updated model
N_c	246	271
N_v	83	83
$N_{\mathrm{D-sel}}$	25	25
$N_{D-\mathrm{val}}$	78	78
LV	14	14
R_c^2	0.95	0.93
RMSEC (g·kg^{-1})	1.45	1.63
R_p^2	0.94	0.93
RMSEP (g·kg^{-1})	1.54	1.58
Bias	0.24	0.25
Slope	0.96	0.96
RMSEP$_{\mathrm{D-sel}}$ (g·kg^{-1})	11.38	1.95 [1]
RMSEP$_{\mathrm{D-val}}$ (g·kg^{-1})	10.45	1.66

[1] RMSECV calculated from the cross-validation results that considered only the prediction errors of the P704 samples in the subset.

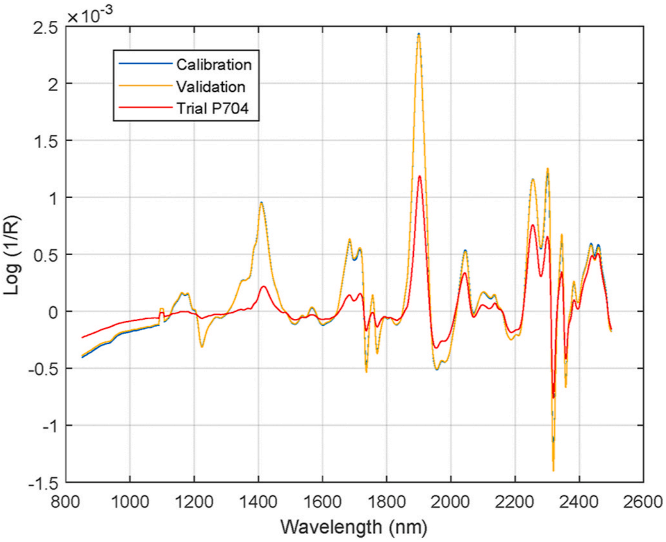

Fig. 1. Mean spectrum (after 1$^{\mathrm{st}}$derivative) of the calibration and validation sets, and that of the new batch of samples.

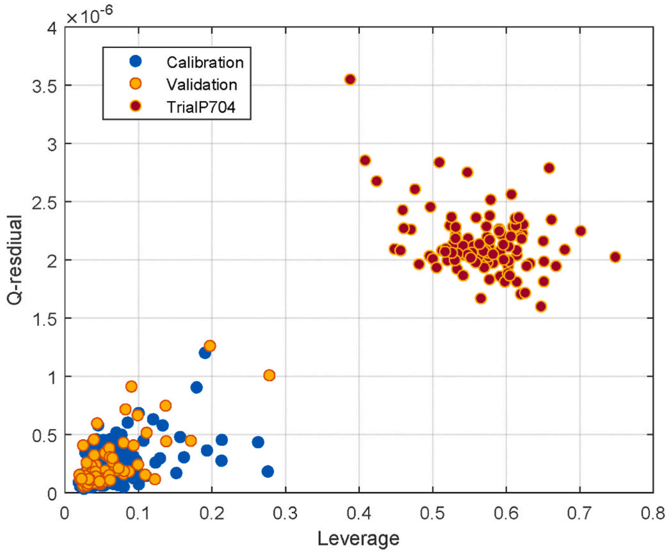

Fig. 2. Q-residuals versus leverage. Calibration samples of the stablished models (blue), validation samples of the stablished models (orange) and samples of the new batch (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

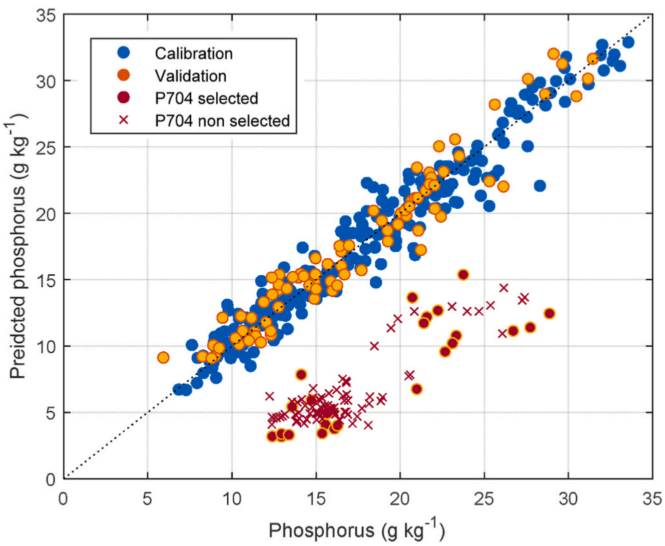

Fig. 3. Predicted vs measured values of phosphorus content with the current model for the calibration set, validation set, the selected samples of Trial P704 and the non-selected samples of Trial P704.

that had been accepted for the current model. Since more digestibility studies of the same type were expected in the future, it was more convenient to update the model than to exceptionally analyze all the new batch of samples with the reference method. Therefore, those same selected samples that had been used to confirm that the predictions errors for this batch were unacceptable, were also used to update the model. The following sections describe the selection of the subset of samples to be analyzed and the validation of the updated model.

3.2. Subset selection

Once the Q-residuals vs leverage plot flagged all 103 samples of trial P704 as outliers, the model scores ($\mathbf{T}$) and the scores of the new trial spectra ($\mathbf{T}_\mathrm{B}$) were submitted to the selection algorithm (section 2.3) to select which samples should be analyzed with the reference method. The algorithm returned the subsets of size $n = 1, \ldots 103$ that maximized Det$\left(\frac{\mathbf{T}_\mathrm{E}^\mathrm{T}\mathbf{T}_\mathrm{E}}{n+N}\right)^{\frac{1}{p}}$. Although subsets with very low n are expected to be useless, they serve to understand the evolution of the optimization criterion when n varies. Fig. 4 shows the value of this determinant against the number of samples in the subset, n. The large increase in the determinant on the left of the graph indicates that the subsets with low n incorporate informative samples that can improve the model. Adding more samples continues to increase the information content of the subset, but the improvement is less and less because the newly selected samples were less unique (that is, their spectra were like the spectra of the already selected samples). A plateau is reached for subsets containing 20 to 30 samples, obtaining the maximum determinant for all subset sizes with the 25-sample subset. For subsets of more than 25 samples, the additional samples did not contribute significant new information per sample and the determinant begins to decrease. Therefore, those samples in the subset of 25 were the most informative and were analyzed with the reference method. They were first used to confirm the need to update the model (as discussed in the previous section) and to update the model.

3.3. Model update and validation

The model was recalculated by adding the 25 selected samples from the trial P704 into the existing training set. The optimal number of LVs obtained by cross-validation was 14, as in the original model. The fact that the inclusion of new sources of spectral variance did not increase the number of LVs means that the variance was distributed over the many LVs of the model. A similar behavior had been observed by Capron et al. [11], who noted that only one out of four models that had been updated required an additional factor while the rest used the same number.

The validation of the updated model was as follows. Commonly, the original data sets are divided into training, validation and test sets that are used to compute the model (training set), select model parameters such as the number of factors or guide wavelength selection (validation set) and verify the actual performance of the final model (test set). When the number of samples is low, the validation set (and sometimes the test set as well) is replaced by some alternative validation method such as cross-validation. In the case of updating a model, the validation of the model with an independent set of samples would require analyzing with the reference method not only those necessary to update the model (as indicated by the selection algorithm) but also a sufficient number to

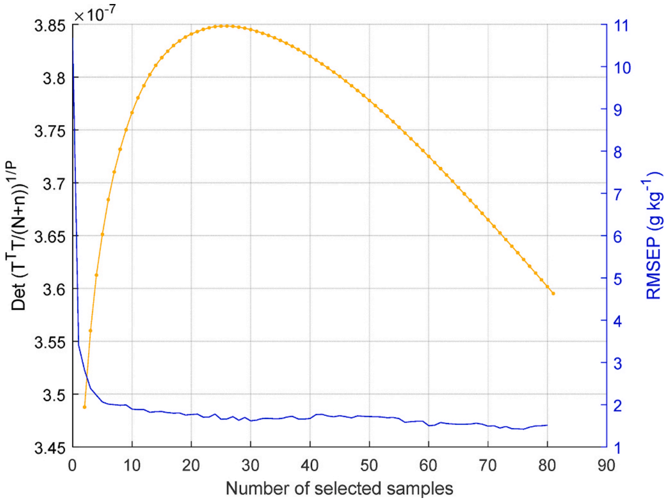

Fig. 4. Number of selected samples used to update the model against the D-criterion. The root mean square error of prediction (RMSEP) of the non-selected samples of trial P704 is also shown.

verify that the model can predict the new samples correctly. However, the analysis of too many samples reduced the benefit of the presented approach and it was decided that no additional samples would be analyzed in addition to those selected by the algorithm. Therefore, validation was carried out by predicting the existing validation set, to confirm that the model maintained the prediction ability for the previous samples, and by cross-validation. The common cross-validation returns an average prediction error over all the samples included in the model (in the form of the root-mean-squared error of cross-validation RMSECV). This obscures the performance of the model for the newly included samples. To focus only on the new samples, a variation of the common RMSECV was calculated from the cross-validation results that considered only the prediction errors of the samples of the new trial. As it can be seen in Table 1, the updated model maintained its previous performance for the samples in the validation set (an RMSEP of 1.58 g kg^{-1} compared with an RMSEP of 1.54 g kg^{-1} before the update) and the prediction error for the new samples was RMSECV = 1.95 g kg^{-1}, much lower than the RMSEP = 11.38 g kg^{-1} obtained before the model update. As discussed, a fairer comparison of prediction errors could be obtained by analyzing extra samples of the new trial. However, this additional effort was not considered to be necessary since the improvement after the update was significant enough to conclude that the update was successful. The updated model was still valid for predicting the original type of samples and also for predicting the rest of the samples of the P704 trial. This finished the model update procedure.

3.4. Performance of the subset selection approach

Exclusively for this work, the 78 not selected samples from the P704 trial were also analyzed with the reference method to study the performance of the presented approach. It should be clear that this is not part of the model update strategy but was necessary to compare the selection strategy with other options.

As it has been shown previously (section 3.1), the need to update the model was discovered from the outlier diagnostics (Fig. 2) and confirmed by the prediction errors of a subset of selected samples from the P704 trial. Fig. 3 shows the predicted phosphorus content with the original model for the selected subset of samples (25 red dots) but also for the unselected samples (78 red crosses) against their reference values. As expected from Fig. 2, the prediction errors were large for all samples in the P704 trial and confirm the conclusions obtained from only the selected samples.

Fig. 4 shows the evolution of the RMSEP of the unselected samples of P704 trial when updated models with an increasing number of optimal samples were used to predict them. The RMSEP remained high when only a few new samples were used to update the model, since the influence of these samples in the calculated model was diluted by the many others in the training set. As expected, the RMSEP decreased as more samples of the P704 trial were included in the model until it reached a stable value of 1.66 g kg^{-1} when the 25 new subset samples were added. This means that the model finally accounted for the spectral variability in the new samples and was able to accurately predict the rest of the samples of the new batch. This model performance correlated well with the information content per sample that was used as a criterion to decide the optimal number of samples. The larger reduction of RMSEP occurred when the increments of the determinant were large which happens with the small size subsets. The RMSEP then kept decreasing as the subset sizes increased while the determinant increased, and the stable RMSEP was reached for the subset with the maximum determinant, at 25 samples. This behavior suggested that the D-optimal criterion helps to select informative samples to update the model.

Fig. 5 shows the predictions of the training samples of the updated model, showing those that were part of the original training set, and those that were added from the selection algorithm. It can be observed that the selected samples, which were predicted with large errors before the update, are now predicted with low errors. This was to be expected, as these samples have now been used in the training step. The most relevant results are those of the unselected samples (red diamonds). As Table 1 shows, these samples had large prediction errors in the original model (RMSEP of 10.45 g kg^{-1}) although this value would never actually be known in the proposed approach because the reference values of these samples would not be known. Now they are predicted well, with an RMSEP of 1.66 g kg^{-1}, similar to the 1.95 g kg^{-1} value obtained by cross-validation of the updated model and lower than the maximum accepted error for this model (2 g kg^{-1}). The prediction of the samples that had not been used for the model update was one of the objectives of the update.

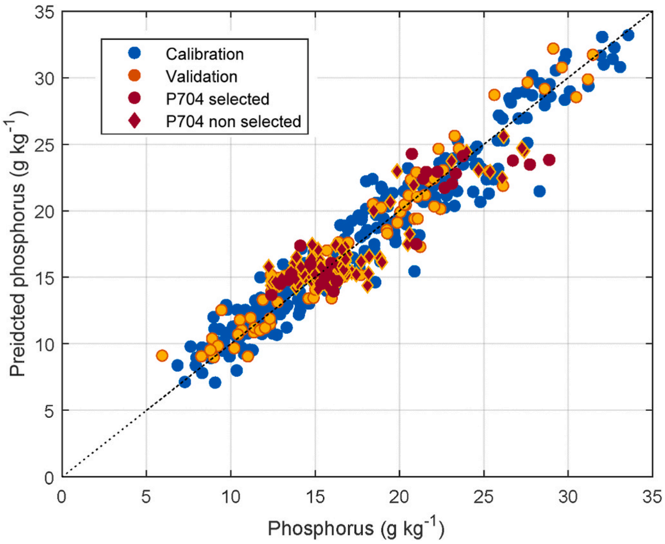

Fig. 5. Predicted vs measured values of phosphorus content with the updated model for the calibration set, validation set, the selected samples of Trial P704 and the non-selected samples of Trial P704.

Fig. 6 compares the effectiveness of the proposed approach with other sample selection methods, namely random selection, the Kennard-Stone algorithm and the selection of the samples with the highest leverage, fixing to 25 the number of samples selected by each method. To select the samples randomly, it must be considered that any selected

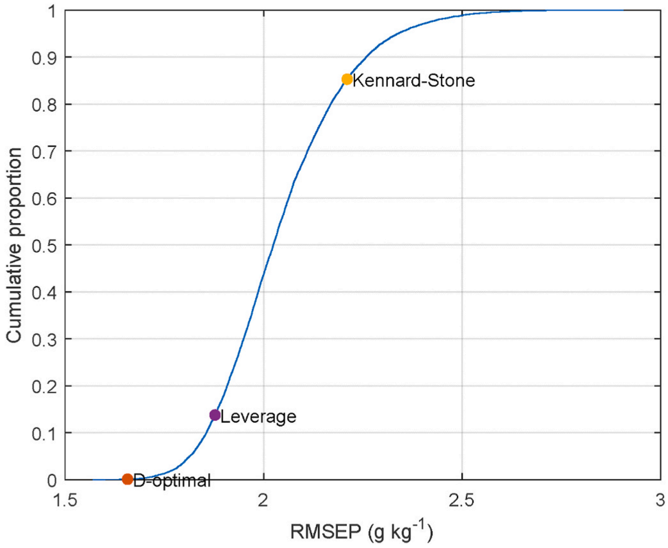

Fig. 6. Cumulative distribution of the RMSEP of 10,000 phosphorus models updated using a subset of samples selected at random. The RMSEP obtained selecting the samples by the D-optimal criterion, Kennard-Stone and sorted leverage are indicated.

subset of 25 samples is out one of the many possible subsets (5.6×10^{23} combinations) that can be created from 103 samples. The performance of these subsets can be ranked to see how one approach compares to the others. For this purpose, 10.000 models were calculated with 25 randomly selected samples from the P704 trial, with the number of LVs selected by cross-validation, and the RMSEP of the unselected samples was plotted as a cumulative distribution in Fig. 6. On the curve one can read the proportion of models that gave an RMSEP less than or equal to a given value. It is seen that random sampling can provide updated models with an RMSEP as low as 1.57 g kg^{-1} but also as high as 2.91 g kg^{-1}. Considering that the acceptable RMSEP for the determination of phosphorus content in the digestion studies is 2 g kg^{-1}, it is seen that the effort of updating a model with 25 random samples can sometimes be useless and can end-up with updated models that predict the remaining samples very poorly (although better than without updating the model). To ensure the efficiency of the experimental effort, a criterion must be sought to guide the selection of the subset so that the RMSEP is as low as possible. The proposed algorithm based on the D-optimality criterion selected a subset whose updated model was better than 99.5% of the randomly subsets selected. The subset selected by sorting the leverage also performed better than most random selections but worse than the D-criterion. In this dataset, the Kennard-Stone algorithm performed the worst. We could not find an explanation for such a bad performance, and the usual performance of the Kennard Stone algorithm found with models of other properties (results not shown) is usually better than the average of random sampling, although still worse than that of D-optimality.

4. Conclusions

A new sample selection algorithm inspired by the D-optimality criterion was successfully used to update a functional PLS model that predicts the phosphorus content of pig faeces from their NIR spectra. The selection algorithm used only the information already available, that is, the complete set of spectra from a new batch. Those spectra had already been measured since the samples were intended to be predicted by the current model. Once outlier detection diagnostics had shown the failure of the model for these samples, Fedorov's algorithm was used to select a subset of the batch that was used to update the model. The new model was validated by cross-validation. The optimal selection of additional samples to be used as a test set to validate the updated model was not considered and is the subject of current research. The results showed that the predictive ability of the updated model was significantly better than the prediction ability of other updated models after selecting the subset by other sample selection methods such as random selection, Kennard-Stone and leverage-based selection.

CRediT authorship contribution statement

Andrés Cruz-Conesa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Joan Ferré:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Itziar Ruisánchez:** Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Anna M. Pérez-Vendrell:** Conceptualization, Investigation, Resources, Data Curation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

A. Cruz-Conesa acknowledges Universitat Rovira i Virgili (Tarragona, Spain) and the Institute of Agrifood Research and Technology (Constantí, Spain) for providing a Marti Franqués Research Fellowship (2019PMF-PIPF-62). The authors thank the various IRTA researchers that made the faeces samples available for this study, and the laboratory personnel for providing the reference values.

References

[1] P.C. Garnsworthy, J. Wiseman, K. Fegeros, Prediction of chemical, nutritive and agronomic characteristics of wheat by near infrared spectroscopy, J. Agric. Sci. 135 (2000) 409–417.

[2] D. Cozzolino, A. Fassio, E. Fernández, E. Restaino, A. La Manna, Measurement of chemical composition in wet whole maize silage by visible and near infrared reflectance spectroscopy, Anim. Feed Sci. Technol. 129 (2006) 329–336.

[3] M.R. Akkaya, Prediction of fatty acid composition of sunflower seeds by near-infrared reflectance spectroscopy, J. Food Sci. Technol. 55 (2018) 2318–2325.

[4] B. Carbas, N. Machado, D. Oppolzer, L. Ferreira, C. Brites, E.A.S. Rosa, A.I.R.N. A. Barros, Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for the determination of nutritional and aninutritional parameters in common beans, Food Chem. 306 (2020), 125509.

[5] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero-Ginel, A. Gómez-Cabrera, Near-infrared reflectance spectroscopy (NIRS) for the mandatory labelling of compound feedingstuffs: chemical composition and open-declaration, Anim. Feed Sci. Technol. 116 (2004) 333–349.

[6] E. Fernández-Ahumada, J.E. Guerrero-Ginel, D. Perez-Marín, A. Garrido-Varo, Near infrared spectroscopy for control of the compound-feed manufacturing process: mixing stage, J. Near Infrared Spectrosc. 16 (2008) 285–290.

[7] L. Paternostre, V. Baeten, B. Ampe, S. Millet, J. De Boever, The usefulness of NIRS calibrations based on feed and feces spectra to predict nutrient content digestibility and net energy of pig feeds, Anim. Feed Sci. Technol. 281 (2021), 115091.

[8] A. Cruz-Conesa, J. Ferré, A.M. Pérez-Vendrell, M.P. Callao, I. Ruisánchez, Use of visible-near infrared spectroscopy to predict nutrient composition of poultry excreta, Anim. Feed Sci. Technol. 283 (2022), 115169.

[9] J.J. Workman Jr., A review of calibration transfer practices and instrument differences in spectroscopy, Appl. Spectrosc. 72 (2018) 340–365.

[10] P. Shahbazikhah, J.H. Kalivas, A consensus modeling approach to update a spectroscopic calibration, Chemometr. Intell. Lab. Syst. 120 (2013) 142–153.

[11] X. Capron, B. Walczak, O. Noord, D. Massart, Selection and weighting of samples in multivariate regression model updating, Chemometr. Intell. Lab. Syst. 76 (2005) 205–214.

[12] D. Bastianelli, L. Bonnal, Y. Jaguelin-Peyraud, J. Noblet, Predicting feed digestibility from NIRS analysis of pig faeces, Animal 9 (2015) 781–786.

[13] F. Westad, F. Marini, Validation of chemometric models – a tutorial, Anal. Chim. Acta 893 (2015) 14–24.

[14] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Transfer of multivariate calibration models: a review, Chemometr. Intell. Lab. Syst. 64 (2002) 181–192.

[15] K. Rajer-Kanduc, J. Zupan, N. Majcen, Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment, Chemometr. Intell. Lab. Syst. 65 (2003) 221–229.

[16] R.H. Galvao, M.C. Araujo, G.E. José, E.C. Silva, T.C. Saldanha, A method for calibration and validation subset partitioning, Talanta 67 (2005) 736–740.

[17] Z.D. Lin, Y.B. Wang, R.J. Wang, L.S. Wang, C.P. Lu, Z.Y. Zhang, L.T. Song, Y. Liu, Improvements of the Vis-NIRS model in the prediction of soil organic matter content using spectral pretreatments, sample selection, and wavelength optimization, J. Appl. Spectrosc. 84 (2017) 529–534.

[18] H.A. D Filho, R.K.H. Galvao, M.C.U. Araújo, E.C. da Silva, T.C. Saldanha, G.E. José, C. Pasquini, I.M. Raimundo Jr., J.J.R. Rohwedder, A strategy for selecting calibration samples for multivariate modelling, Chemometr. Intell. Lab. Syst. 72 (2004) 83–91.

[19] Z. He, M. Li, Z. Ma, Design of a reference value-based sample-selection method and evaluation of its prediction capability, Chemometr. Intell. Lab. Syst. 148 (2015) 72–75.

[20] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.

[21] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.

[22] G. Puchwein, Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra, Anal. Chem. 60 (1988) 569–573.

[23] J.S. Shenk, M.O. Westerhaus, Population, definition, sample selection, and calibration procedures for near-infrared reflectance spectroscopy, Crop Sci. 31 (1991) 469–474.

[24] B. Xu, Z. Wu, Z. Lin, C. Sui, X. Shi, Y. Qiao, NIR analysis for batch process of etahanol precipitation coupled with a new calibration model updating strategy, Anal. Chim. Acta 720 (2012) 22–28.

[25] W. Chen, Y. Yun, M. Wen, H. Lu, Z. Zhang, Y. Liang, Representative subset selection and outlier detection via isolation forest, Anal. Methods 8 (2016) 7225–7231.

[26] P.F. Aguiar, B. Bourguignon, M. Khots, D. Massart, R. Phan-Than-Luu, D-optimal designs, Chemometr. Intell. Lab. Syst. 30 (1995) 199–210.

[27] J. Ferré, F.X. Rius, Constructing D-optimal designs from a list of candidate samples, Trends Anal. Chem. 16 (1997) 70–73.

[28] J. Ferré, F.X. Rius, Selection of the best calibration sample subset for multivariate regression, Anal. Chem. 68 (1996) 1565–1571.

[29] Official Methods of Analysis of AOAC Internacional, twentieth ed., AOAC Internacional, Gaithersburg, MD, 2016.

[30] A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.

[31] V.V. Fedorov, in: W.J. Studden, E.M. Klimko (Eds.), Theory of Optimal Experiments, Translated, Academic Press, New York, 1972.

[32] A. Broudiscou, R. Leardi, R. Phan-Tan-Luu, Genetic algorithm as a tool for selection of D-optimal design, Chemometr. Intell. Lab. Syst. 35 (1996) 105–116.

[33] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Artificial neural networks in classification of NIR spectral data: design of the training set, Chemometr. Intell. Lab. Syst. 33 (1996) 35–46.

[34] O.Y. Rodionova, A.L. Pomerantsev, Subset selection strategy, J. Chemom. 22 (2008) 674–685.

[35] ASTM E1655-17. Standard practices for infrared multivariate quantitative analysis; American Society for Testing and Materials. ASTM International, West Conshohocken, PA.

[36] S.K. Setarehdan, J.J. Soraghan, D. Littlejohn, D.A. Sadler, Maintenance of a calibration model for near infrared spectrometry by a combined principal component analysis-partial least squares approach, Anal. Chim. Acta 452 (2002) 35–45.

[37] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1984.