PROSAC as a selection tool for SO-PLS regression: a strategy for multi-block data fusion

Authors

Jose A. Diaz-Olivares[1a], Ryad Bendoula[b], Wouter Saeys[c], Maxime Ryckewaert[b,d], Ines Adriaens[a,e], Xinyue Fu[a], Matti Pastell[f], Jean-Michel Roger[b,d], Ben Aernouts[a]

[a] KU Leuven, Department of Biosystems, Division of Animal and Human Health Engineering, Campus Geel, Kleinhoefstraat 4, 2440 Geel, Belgium

[b] ITAP, Univ. Montpellier, INRAE, Institute Agro, Montpellier, France

[c] KU Leuven, Department of Biosystems, MeBioS unit, Kasteelpark Arenberg 30, 3001 Leuven, Belgium

[d] ChemHouse Research Group, Montpellier, France

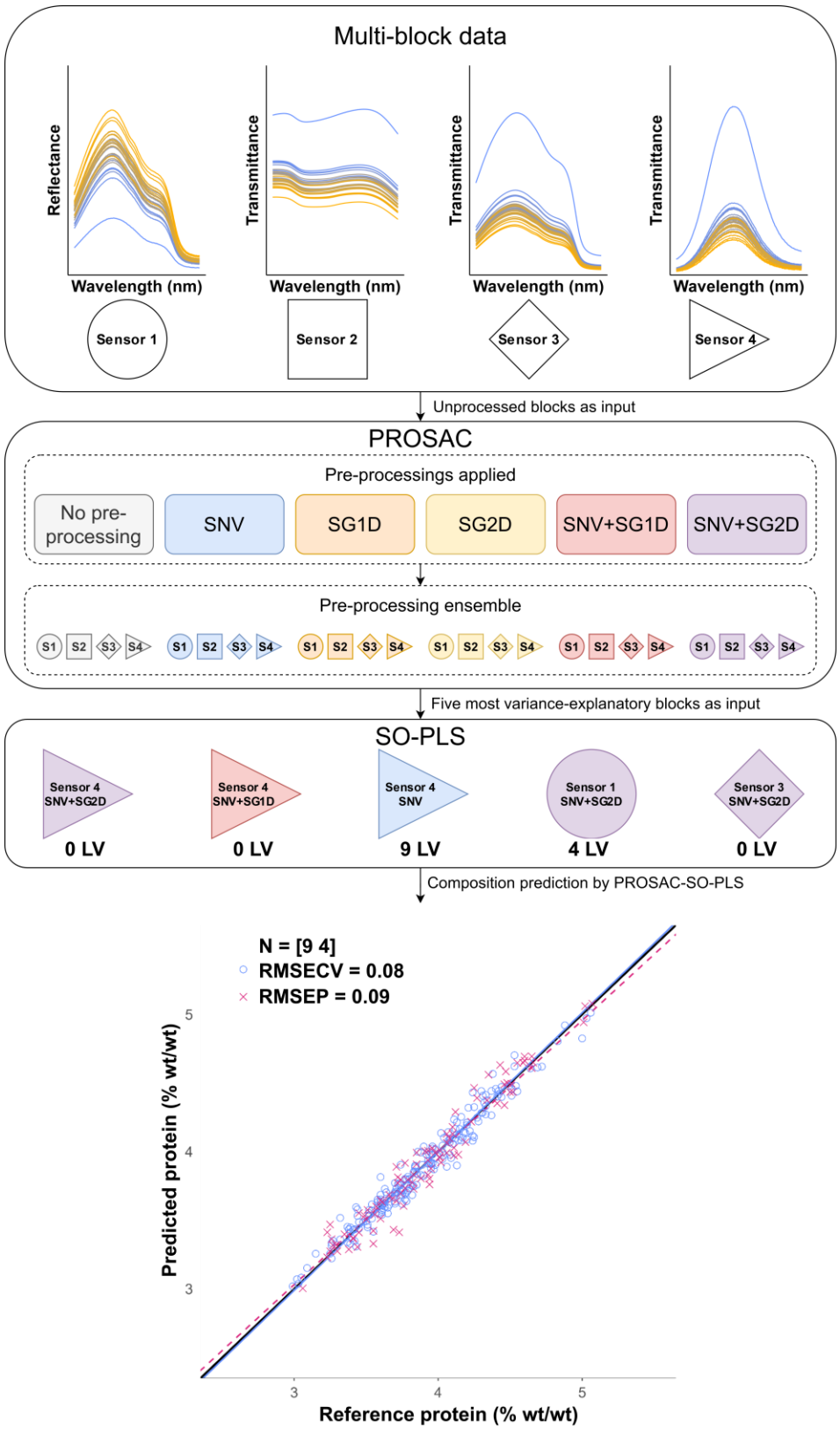
[e] Department of Data Analysis and Mathematical Modelling, Division BioVism, Campus Coupure, Coupure Links 653, 9000 Ghent, Belgium

[f] Production Systems, Natural Resources Institute Finland (Luke), Latokartanonkaari 9, 00790 Helsinki, Finland.

[1] Corresponding author: Jose A. Diaz-Olivares, KU Leuven, Department of Biosystems, Division of Animal and Human Health Engineering, Campus Geel, Kleinhoefstraat 4, 2440 Geel, Belgium, +32 (0)14 72 13 64, jose.diaz@kuleuven.be

---

Graphical Abstract



Graphical abstract. *Summary of the PROSAC-SO-PLS methodology. Multi-block spectral data derived from multiple sources is subjected to an ensemble of pre-processing techniques within the PROSAC framework. The five most variance-explanatory blocks identified by PROSAC are subsequently utilized as input variables for SO-PLS, facilitating the construction of a composition prediction model that optimizes block utilization.*

Abstract

Spectral data from multiple sources can be integrated into multi-block fusion chemometric models, such as sequentially orthogonalized partial-least squares (SO-PLS), to improve the prediction of sample quality features. Pre-processing techniques are often applied to mitigate extraneous variability, unrelated to the response variables. However, the selection of suitable pre-processing methods and identification of informative data blocks becomes increasingly complex and time-consuming when dealing with a large number of blocks. The problem addressed in this work is the efficient pre-processing, selection and ordering of data blocks for targeted applications in SO-PLS.

We introduce the PROSAC-SO-PLS methodology, which employs pre-processing ensembles with response-oriented sequential alternation calibration (PROSAC). This approach identifies the best pre-processed data blocks and their sequential order for specific SO-PLS applications. The method uses a stepwise forward selection strategy, facilitated by the rapid Gram-Schmidt process, to prioritize blocks based on their effectiveness in minimizing prediction error, as indicated by the lowest prediction residuals. To validate the efficacy of our approach, we showcase the outcomes of three empirical near-infrared (NIR) datasets. Comparative analyses were performed against partial-least-squares (PLS) regressions on single-block pre-processed datasets and a methodology relying solely on PROSAC. The PROSAC-SO-PLS approach consistently outperformed these methods, yielding significantly lower prediction errors. This has been

evidenced by a reduction in the root-mean-squared error of prediction (RMSEP) ranging from 5 to 25% across seven out of the eight response variables analyzed.

The PROSAC-SO-PLS methodology offers a versatile and efficient technique for ensemble pre-processing in NIR data modeling. It enables the use of SO-PLS minimizing concerns about pre-processing sequence or block order and effectively manages a large number of data blocks. This innovation significantly streamlines the data pre-processing and model-building processes, enhancing the accuracy and efficiency of chemometric models.

Keywords

Multi-block ; Chemometrics; Pre-processing; Spectroscopy; Data fusion; NIR;

1 Introduction

The landscape of analytical chemistry has been transformed by the proliferation of sensor devices and instrumental techniques, coupled with the use of multivariate data analysis [1]. Collecting multi-source data from identical samples, for instance through the simultaneous use of multiple spectrometers operating across different wavelength ranges [2–5], facilitates the creation of multi-block datasets. The fusion of these data enhances the analysis of the variables of interest, improving both interpretation and prediction capabilities compared to a single-block data approach [6], and uncovering unique and common variations across data sources [7].

In the field of analytical techniques, near-infrared (NIR) spectroscopy stands out for its cost-effectiveness [8], minimal labor and sample preparation, and the absence of chemical reagents [9]. In many applications, miniature NIR spectrometers have replaced classic benchmark systems, allowing for specialized, multi-device deployments [10]. This has led to the widespread application of multi-source data fusion combined with NIR spectroscopy for rapid, non-destructive evaluation of physicochemical properties [11–13].

When capturing multi-source variability, conventional single-block chemometric methods such as partial-least-squares regression (PLS) [14] and principal component analysis (PCA) [15] can be insufficient [16]. Their primary limitation in this context is the inability to effectively handle complex inter-block relationships inherent in multi-source datasets. To address this limitation, a wide array of low-level model-based data fusion approaches have been proposed for extracting unique features and variability from multiple data sources [17,18]. These data fusion approaches have specific considerations for the multi-block nature of the data while still allowing the influence of each source to be identifiable in subsequent model building [19].

A particular data fusion method is sequential and orthogonalized partial-least-squares (SO-PLS) [20], a multi-block extension of PLS. It sequentially integrates data blocks, quantitatively evaluating the

contribution of each predictor block to the predictive capability of the model, while avoiding redundancies through sequential orthogonalization [21]. Moreover, SO-PLS is scale-invariant and can handle blocks of varying sizes without spurious bias. Comparative studies indicate that it outperforms the prediction accuracy of other state-of-the-art multi-block algorithms [22].

The order in which blocks are input into SO-PLS significantly influences both model coefficients and complexity [23], complicating interpretation and requiring expert knowledge to determine optimal block order [24]. While some SO-PLS-based methods mitigate this issue (Campos et al., 2018), pre-processing techniques should also be considered [26], as they are often essential to minimize non-linearity and scattering effects in spectral data while enhancing the signal-to-noise ratio [27]. SO-PLS-based methods like SPORT [28] enable ensemble pre-processing for spectral data but with high computational costs. Moreover, the exhaustive search for optimal latent variable combinations becomes more complex as different pre-processing steps increase the number of data blocks, which not only increases computational demands but also limits the exploration of diverse pre-processing options.

To address the challenge of the time and computing resource-intensive nature of managing a high number of data blocks, the response-oriented sequential alternation (ROSA) algorithm serves as a fast multi-block extension of PLS modeling [29]. ROSA employs a rapid Gram-Schmidt process to circumvent the computational intensity of deflations and uses stepwise selection to select blocks that result in the lowest prediction residuals, thus optimizing model performance by minimizing prediction error.

This procedure enables the handling of a large number of blocks efficiently. As a result, ROSA can be used as a calibration tool for large pre-processing ensembles, a combination known as pre-processing ensembles with response-oriented sequential alternation calibration or PROSAC [30]. With its scale and order invariance and lack of spurious bias, PROSAC is well-suited for evaluating all relevant pre-processings in a given multi-block dataset, assigning equal priority to each through its parallel approach.

Despite its many advantages, PROSAC presents certain limitations. The algorithm, based on forward stepwise selection, dictates that the initial choice of a certain block can influence the dataset in subsequent deflation phases, potentially affecting later selections and causing the optimization to settle at local minima [31]. Moreover, the stepwise selection of the best predictor block inherently increases the risk of overfitting. This is due to the algorithm relying on both covariance, used in score building, and correlation, guiding block selection. The risk is further increased when the process iterates over the same dataset multiple times. Additionally, PROSAC may converge in models with a large number of included blocks, becoming counterproductive especially when employed as a decision-making tool in the development of experiments or sensor systems. In such contexts, each block might represent a different data source, for instance, a distinct miniature spectrometer, thereby complicating the physical setup. This physical complexity can make it difficult to utilize all selected data sources effectively, which in turn could compromise the accuracy of compositional predictions in the applied PROSAC model.

To address the identified limitations of PROSAC, we propose to leverage it as a generalized variable selection method in which the multiple blocks emulate variables. Initially, PROSAC applies a pre-processed ensemble to a raw multi-block dataset, selecting up to a maximum number of blocks M (as outlined in the original PROSAC publication [30]) that best explain the variations in the response variables. This initial limit on M is guided by efficiency and the need to manage the computational load. Subsequently, a more limited and ordered selection of no more than K of the most explanatory blocks identified by PROSAC is implemented in SO-PLS. This intentional restriction to a maximum of K blocks serves to construct an accurate prediction model while minimizing the computational costs and time use inherent in SO-PLS. This approach is referred to as the PROSAC-SO-PLS methodology. To the best of our knowledge, no research has been conducted on exploring the application of PROSAC to determine the ideal blocks for a more targeted implementation like SO-PLS.

This study presents a comprehensive evaluation of the PROSAC-SO-PLS methodology and compares it with current state-of-the-art approaches, including a single-block approach and a standalone PROSAC. The effectiveness of the methodology is demonstrated through three empirical datasets pertinent to multi-block applications within agrifood research.

2 Materials and methods

2.1 PROSAC implementation

To build prediction models using a multi-block approach, raw multi-block data must first be extracted from the dataset, followed by the application of pre-processing methods. Each individual block consists of an $X_{n\times p_j}$ spectral array, where n represents the sample count and p_j stands for the spectral variables for that block. It is important to note that each block can have a distinct number of variables p_j, accommodating different spectral sources or pre-processing methods that might alter the variable count. A dataset has j different $X_{n\times p_j}$ arrays, or blocks, with j determined by the count of spectral sources within that set. Consequently, each dataset can be summarized with a unique raw multi-block structure $Z = [X_1; X_2; X_3; X_4 \dots X_j]$, paired to a response array $Y_{n\times k}$. In this context, k represents the total number of measured responses.

The pre-processings applied to the different spectral arrays in this study were adapted from general methods applied in the original PROSAC approach by Mishra et al. [30]: no processing, first-order Savitzky-Golay derivative (SG1D), second-order Savitzky-Golay derivative (SG2D), standard normal variates (SNV) weighting, and combinations of SNV with SG1D or SG2D. These resulted in six multi-block structures, presented as $[Z_1; Z_2; Z_3; Z_4; Z_5; Z_6]$, with each structure corresponding to a specific pre-processing method applied to the raw multi-block structure (Z). Consequently, the final input to the PROSAC algorithm consists of a multi-block ensemble of $6 \times j$ different $X_{n\times p_j}$ blocks. Changing the order in which the pre-processing blocks serve as inputs in PROSAC was not considered, as it is inherently order-independent in its processing and analysis [30].

The PROSAC algorithm processes the obtained multi-block ensemble, leveraging the heuristic of ROSA to iteratively select blocks for building the prediction model through a competition mechanism. In each

iteration, the block with the smallest prediction residual is chosen, allowing for their repeated use to extract complementary information, and defining the model component for that stage. To determine the optimal number of blocks necessary for model construction, ROSA relies on the minimum root-mean-squared error of calibration (RMSEC). Extending the ROSA approach, PROSAC conducts a 10-fold cross-validation on the calibration set using random groupings. This approach effectively reduces the risk of overfitting, enhancing model generalizability.

This process is repeated 10 times to identify the combination of blocks that yield the minimum root-mean-squared error of cross-validation (RMSECV), with PROSAC potentially selecting the same block more than once, up to a maximum of 50 blocks ($M = 50$). To mitigate overfit, this application of PROSAC selects the smallest number of blocks that ensures an RMSECV statistically indistinguishable from the minimum achievable RMSECV when more blocks are used. One-sided paired t-tests on the squared residuals are conducted to confirm this selection, using a significance level of 0.05 [32]. This approach is applied separately for each individual response variable in the dataset, yielding a corresponding prediction model.

2.2 Integrating SO-PLS with PROSAC block selection strategy (PROSAC-SO-PLS)

SO-PLS sequentially incorporates input blocks, enhancing the predictive accuracy of the model with each addition. The process begins with a PLS regression (PLSR) between the first input block, identified as the most representative in capturing the variation in the response variable, and the response variable itself. Each subsequent block is orthogonalized against the scores from the preceding PLSR and fitted to its prediction residuals, a step executed one block at a time.

Building on the previously described PROSAC methodology, the first K blocks selected through the iterative process of PROSAC are used as inputs for SO-PLS. The current approach has limited K to five blocks, aiming to optimize model prediction capabilities while efficiently managing computational resources. This choice is substantiated by a detailed analysis later in the text, where the computational

feasibility of using up to five blocks is contrasted with the exponential increase in resource demand for additional blocks. In instances where PROSAC selects repeated blocks to capture complementary information, the next unique block is chosen as a replacement for the SO-PLS input. This is because SO-PLS, due to its inherent orthogonalization, does not extract new information from repeated blocks. By limiting the selection of input blocks, the computational cost associated with SO-PLS decreases. Moreover, since these blocks are ordered by explained variance, the critical dependency of SO-PLS on block order is effectively managed.

The selection of the appropriate number of latent variables for each block employs a repeated 10-fold cross-validation with random groups on the calibration set, repeated 10 times. This approach facilitates a precise estimation of the necessary components for each block by optimizing the latent variables individually in each PLS regression. This strategy, as proposed by [20], is adopted to develop cross-validated SO-PLS models using all possible combinations of latent variables, with a maximum of 20 latent variables per block.

To select the optimal model complexity, a Måge plot was used, which visually represents the prediction error for each combination of individual latent variables and block combinations as a function of the total number of components [20]. This plot aids in identifying the configuration that minimizes RMSECV. Although effective, this approach can increase the number of parameters, raising the risk of overfitting and requiring extensive test set validation [33]. To mitigate this, a parsimony-guided adjustment is applied, selecting the simplest configuration that yields an RMSECV statistically indistinguishable from the minimum, as verified by one-sided paired t-tests on the squared residuals ($\alpha = 0.05$).

If the number of latent variables is equal to zero for any of the initial five input blocks, subsequent iterations evaluate the inclusion of successive PROSAC blocks as replacements. The final model configuration is determined either when the inclusion of new blocks ceases to lower the RMSECV in a

statistically significant manner verified by one-sided paired t-tests on the squared residuals (α = 0.05), or when no additional predictor blocks are available for inclusion.

2.3 Experimental datasets

Three distinct experimental datasets were employed to illustrate and validate the PROSAC-SO-PLS methodology, each undergoing spectral normalization using dark and white reference measurements. The first dataset [34], hereafter referred to as the SRS-milk dataset, features spatially-resolved spectroscopy (SRS) reflectance measurements of raw milk in the LW-NIR region (960 to 1690 nm), measured by a 1.7-256 Plane Grating Spectrometer (Carl Zeiss, Jena, Germany). This SRS implementation used two optical fibers, each housed in a metal ferrule dipped at least 25 mm below the surface of the raw milk samples. One fiber illuminated the sample while the other fiber detected the light that interacted with and was reflected by the sample. The detection fiber traversed a horizontal path to capture SRS spectra at 30 equidistant illumination-to-detection distances, ranging from 1.1 to 4 mm. For performance assessment, 186 raw milk samples were measured with the SRS setup, and their fat, protein, and lactose content was determined with the reference methodology [35]. A dynamic range correction was applied on the LW-NIR SRS spectra to compensate for the exponential-like decrease in signal intensity and suboptimal signal-to-noise ratio caused by increasing illumination-to-detection distances in SRS measurements, as demonstrated by Diaz-Olivares et al. [34].

The second dataset [36], named the miniS-milk dataset, contains NIR spectral data measured from 299 raw milk samples using four different NIRONE miniature spectrometers (Spectral Engines, Steinbach, Germany) with complementary wavelength ranges: T14 (NIRONE 1.4; 1100 to 1400 nm, measuring in transmittance mode), R20 and T20 (NIRONE 2.0; 1550 to 1950 nm, two units in reflectance and transmittance mode, respectively) and T25 (NIRONE 2.5; 2000 to 2450 nm, transmittance mode). Fat, protein, and lactose content of the samples was determined with the reference methodology [35].

Finally, for the third experimental dataset [37], referred to as the miniS-sugarcane dataset, reflectance NIR spectral data were gathered from 60 sugarcane samples with a variety of miniature spectrometers, including the F750 (Felix Instrument, Camas, USA; 450 to 1140 nm), SCIO (Consumer Physics, Hod Hasharon, Israel; 740 to 1070 nm), NIRscan (DLP NIRscan Nano; Texas Instruments Inc., Dallas, USA; 901 to 1701 nm), NIR1.7k and NIR2.2k (µNIR1700 and µNIR2200; Viavi, Chandler, USA; 908 to 1676 nm and 1150 to 2150 nm), and NIRONE 2.2 (1750 to 2150 nm). Each of these spectrometers, featuring overlapping wavelength ranges, was employed to measure the complete set of sugarcane samples. Reference measurements for crude protein (CP) content were derived from the total nitrogen content (Nt) measured by the Kjeldahl method (European Commission, 2009), with the relationship $CP = 6.25 * Nt$, while total sugars (TS) were determined by the modified Luff-Schoorl method [39].

Each of the three datasets was split into a calibration set comprising roughly two-thirds of the samples, reserving the remaining one-third of the samples for the test set, employing the Duplex algorithm with Mahalanobis distance [40] to partition the datasets based on the reference response variables. All composition reference values and predictions, unless specified otherwise, are expressed in weight/weight (wt/wt) percentages. Specifically for the miniS-sugarcane dataset, the units are % wt/wt but are based solely on the dry matter content of the sugarcane.

2.4 Development and validation of single-block prediction models

For the three experimental datasets, single-block PLSR prediction models were developed for each spectral source to benchmark against PROSAC and the PROSAC-SO-PLS methodology. While PROSAC-SO-PLS employed a consistent preprocessing ensemble across datasets (none, SG1D, SG2D, SNV, SNV+SG1D, SNV+SG2D) as indicated in Section 2.1, the single-block models utilized fixed dataset-specific preprocessing methods based on prior studies. In the SRS-milk and miniS-milk datasets, spectra were pre-processed using a fixed combination of SNV and Savitzky-Golay derivatives. The derivatives windows were

tailored to the unique absorption characteristics of milk components: 15 wavelengths (±40 nm) for fat and protein using SG1D, and 19 wavelengths (±50 nm) for lactose using SG2D, as found effective in previous studies [41]. For the miniS-sugarcane dataset, the preprocessing adhered to Ryckewaert et al. (2022), using SNV with Savitzky-Golay derivatives but varying window sizes individually adapted to the characteristics of each miniature spectrometer used in this study. Individual PLSR models were then constructed for each response variable and each block across all datasets, with a maximum of 20 latent variables. The specific pre-processings employed for each response variable are detailed in Table 2 in the Results section. Mean centering was applied before PLSR model construction in all instances.

To assess model complexity and performance, each calibration set underwent a 10-fold cross-validation process with random groups of equal size, repeated and randomized 100 times. This evaluation focused on minimizing the RMSECV to select the optimal number of latent variables. Within the PLSR of each single block, the minimal number of latent variables was chosen for which the RMSECV was not statistically different from the minimum RMSECV. This selection was confirmed through one-sided paired t-tests on the squared residuals ($\alpha = 0.05$).

After determining the number of latent variables for each block, predictions for sample compositions were made using the PLSR models on the test sets, calculating the corresponding prediction residuals and the derived root-mean-square error of prediction (RMSEP). For each dataset, the single block yielding the lowest RMSEP was identified as the most accurate for composition prediction and served as a benchmark for comparing with the multi-block methods.

2.5 Performance comparison

The PROSAC (section 2.1) and PROSAC-SO-PLS (section 2.1 followed by section 2.2) methods were applied to the calibration set. Once calibrated and built, each PROSAC and SO-PLS model was then applied to the corresponding test set to calculate the corresponding residuals and determine the respective RMSEP

values. This was done for each individual response variable in the three different datasets. Next, the predictive performances of three approaches (single-block PLSR using optimally pre-processed data, standalone PROSAC, and the combined PROSAC-SO-PLS methodology) were compared.

The comparative analysis evaluated key metrics such as RMSECV, RMSEP, and the number of blocks (N) utilized. To assess the effectiveness of the multi-block methodologies, a paired two-way analysis of variance (ANOVA) was conducted on the squared residuals of the test samples, considering the model type as a three-level factor (single-block, PROSAC, or PROSAC-SO-PLS) and the sample number as a random factor. Only when a significant effect of the correction was detected by the ANOVA procedure ($\alpha = 0.05$), the approaches were compared mutually with a Tukey's HSD multiple comparisons test ($\alpha = 0.05$).

All prediction models were developed and validated using a custom chemometrics toolbox in MATLAB version 2021a (Mathworks, Natick, USA). The specific codes used in the current study are referenced in the data availability section. Computational cost analysis were conducted on a Microsoft Windows 10 Pro OS, utilizing a system equipped with a 4.7 GHz 12-core Ryzen processor (AMD, Santa Clara, USA) and 64 GB of RAM.

3 Results and discussion

3.1 Data overview

Following the Duplex approach, the SRS-milk dataset was divided into 120 samples for calibration and 66 for testing. The miniS-milk dataset was split into a calibration group of 205 samples and a test group of 94 samples, while the miniS-sugarcane dataset was split into 40 calibration samples and 20 test samples. In the SRS-milk dataset, regions between 1360 and 1500 nm were excluded from subsequent analysis due to the diminished SRS signals resulting from water absorption. Additionally, the spectral extremities from 1680 to 1690 nm were eliminated owing to the reduced sensitivity of the spectrometer in these wavelength ranges [34].

Table 1 presents the descriptive statistics and correlations for the reference response variables of the calibration and test sets. A two-sample t-test (α = 0.05) confirmed no significant difference between sets, indicating that the data splits are representative of their corresponding dataset, a critical aspect for robust model development [43]. For both the SRS-milk and miniS-milk datasets, the composition and variability are consistent with other raw milk datasets [44] and findings from milk recording programs [45]. Similarly, the CP and TS contents in the miniS-sugarcane dataset correspond to values reported in other sugarcane studies [46].

Table 1. *Descriptive statistics and Pearson correlations of the predicted response variables in the calibration and test sets across all datasets.*

Dataset	Resp. Var.	Calibration						Test					
		Basic statistics (% wt/wt)				Pearson corr.		Basic statistics (% wt/wt)				Pearson corr.	
		Mean	SD	Min	Max	Comp. #2	Comp. #3	Mean	SD	Min	Max	Comp. #2	Comp. #3
SRS-milk	Fat	4.04	0.58	2.49	5.34	0.41	-0.17	3.99	0.66	2.46	5.46	0.46	-0.21

	Prot.	3.37	0.25	2.55	3.98	1	-0.19	3.37	0.28	2.74	4.03	1	-0.24
	Lact.	4.69	0.13	4.38	5.07	-	1	4.71	0.15	4.31	5.05	-	1
miniS-	Fat	4.66	1.01	1.76	7.62	0.40	-0.33	4.84	1.29	1.71	7.70	0.37	-0.47
milk	Prot.	3.38	0.40	2.99	5.06	1	-0.26	3.93	0.45	3.06	5.07	1	-0.48
	Lact.	4.63	0.15	4.18	5.10	-	1	4.60	0.16	4.09	4.99	-	1
miniS-	CP	3.10	2.19	1.04	9.60	-0.64	-	3.03	1.80	0.90	6.62	1	-
sugarcane	TS	23.81	17.12	1.15	51.01	1	-	23.42	17.33	2.11	50.42	-0.65	-

"Comp. #2" and "Comp. #3" denote Pearson correlations within datasets. A dash (-) indicates correlation is either nonexistent or previously stated. For miniS-sugarcane, units are % wt/wt based on sugarcane dry matter only.

Figure 1 depicts the normalized SRS reflectance spectra after dynamic range correction for all 186 milk samples across three equidistant measuring points (1.1, 2.5, and 4 mm) from the total 30 illumination-to-detection distances (SRS-milk). Figure 2 illustrates the normalized miniS-milk dataset with measurements taken by four miniature spectrometers. Both datasets show significant decreases in reflectance and transmittance at specific wavelengths (970, 1200, 1450, 1940, and above 2400 nm), attributed to high light absorption by water. Fat globules contribute to the back-scattering of a considerable portion of the light towards the light source, which leads to elevated reflectance and reduced transmittance in milk samples with higher fat content [45].

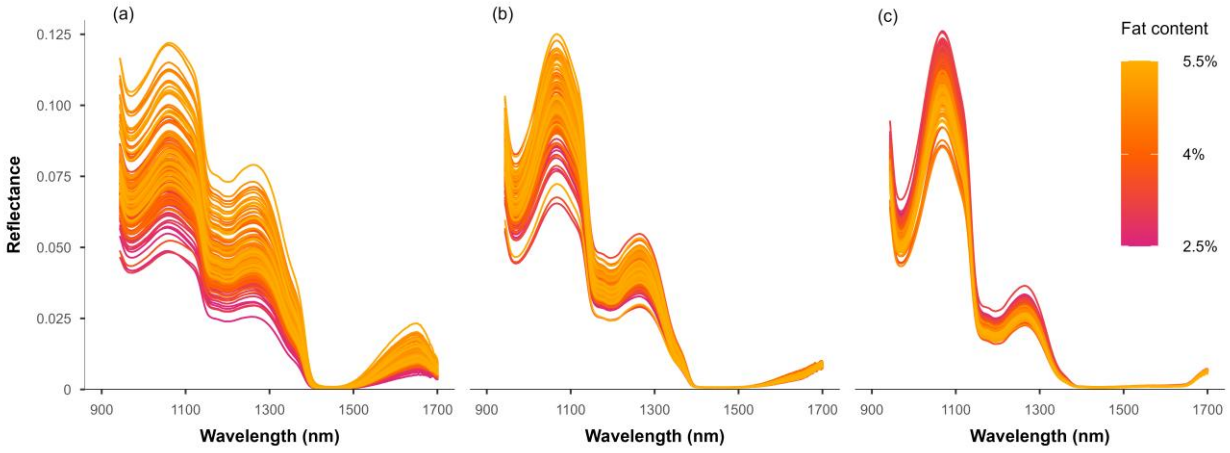


Figure 1. *Normalized LW-NIR SRS-milk reflectance spectra for 186 milk samples, displayed at three illumination-to-detection distances: (a) 1.1, (b) 2.5 and (c) 4 mm. Yellow hues indicate higher fat content; magenta indicates lower levels.*

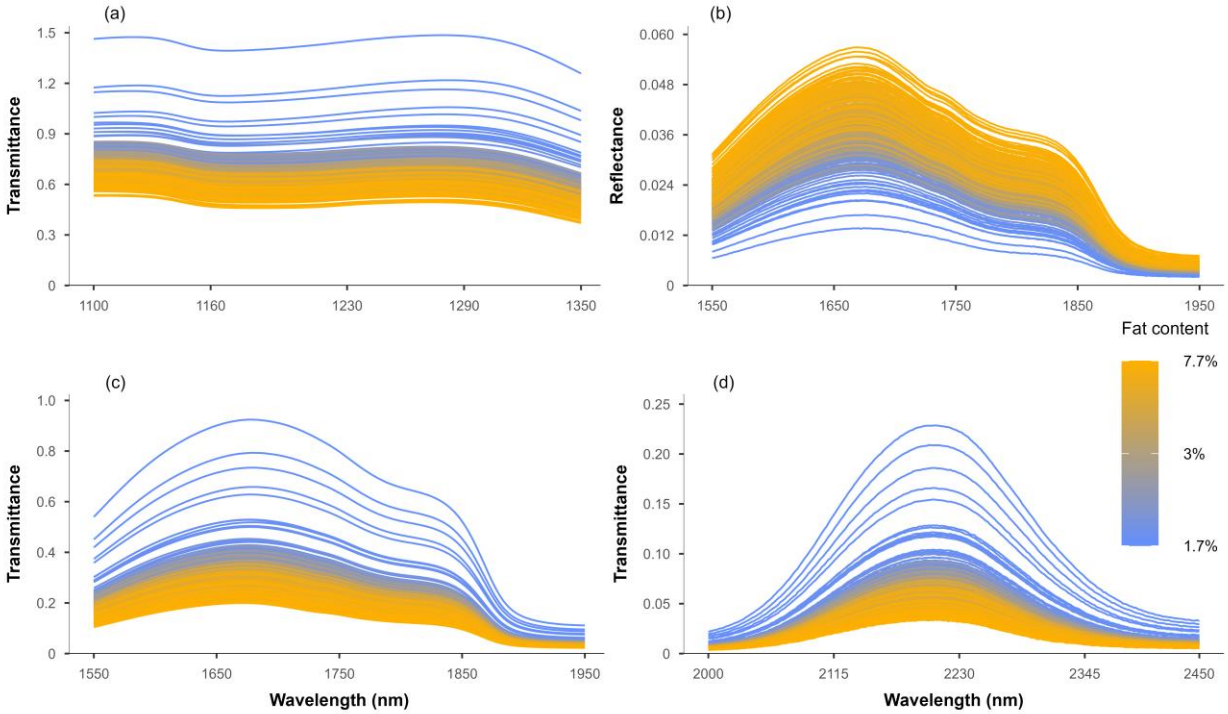


Figure 2. *Normalized miniS-milk spectra from 299 samples, acquired via multiple miniature spectrometer devices: (a) T14, (b) R20, (c) T20 and (d) T25. Yellow and blue hues represent elevated and reduced fat content, respectively.*

Figure 3 displays the NIR spectra captured from 60 sugarcane samples. These spectra highlight the pronounced influence of absorption by water molecules, also found in the previous datasets. Additionally, the reduced reflectance values at 670 nm and 1200 nm can be attributed to absorption by chlorophyll and sucrose, respectively [47]. An increase in sugar content correlates with higher overall NIR absorbance and thus a lower reflectance [48].

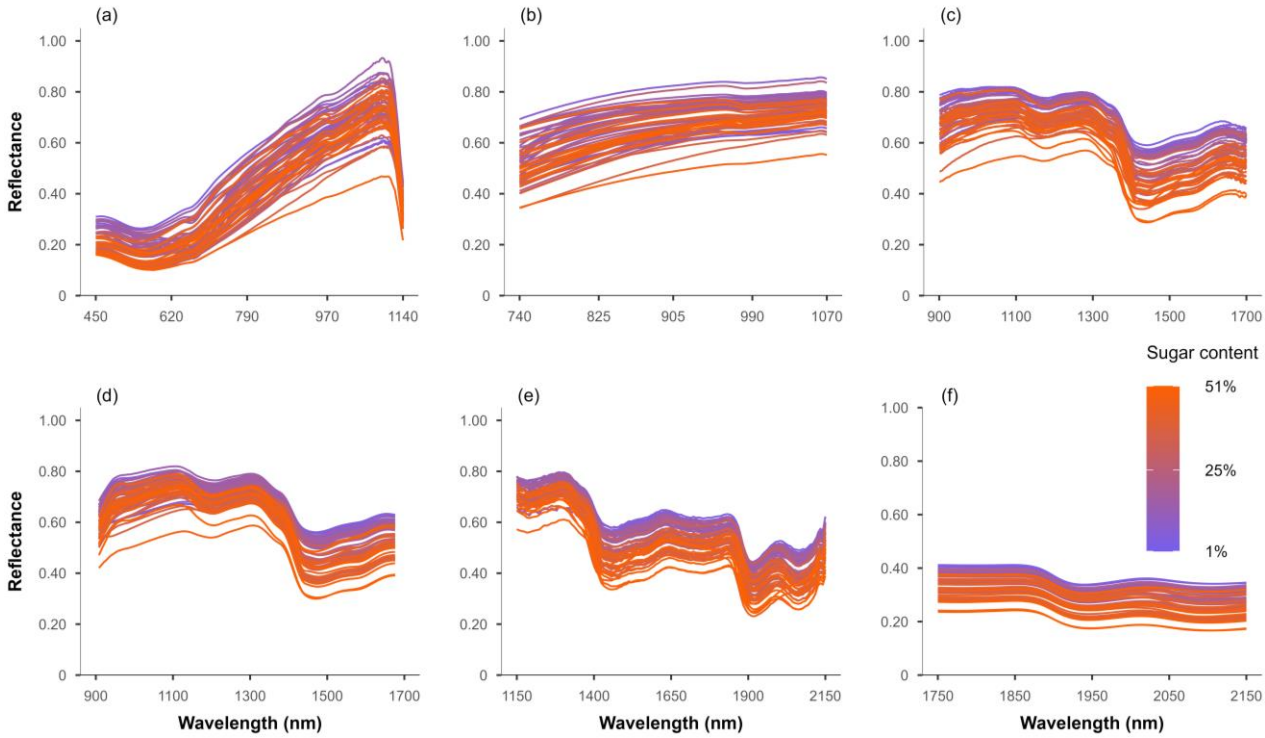


Figure 3. *Normalized miniS-sugarcane reflectance spectra for 60 sugarcane samples, collected using different miniature spectrometers: (a) F750, (b) SCIO, (c) NIRscan, (d) NIR1.7k, (e) NIR2.2k and (f) NIRONE 2.2. Orange-to-purple hues indicate higher to lower sugar content.*

3.2 Evaluation of the single-block prediction models

The effectiveness of the standalone PROSAC and the combined PROSAC-SO-PLS approach was assessed using the performance of single-block PLSR modeling on the three datasets as a benchmark. Table 2 presents the predictive accuracy of the most effective single block for each response variable in each dataset. Performance metrics include the RMSECV, RMSEP and number of latent variables.

Table 2. Prediction performance statistics for the best single-block prediction model within a given response variable, for all datasets.

Dataset	Response variable	Block	Pre-processing	LV	RMSECV (% wt/wt)	RMSEP (% wt/wt)
SRS-milk	Fat	2.3 mm	SNV+SG1D(15)	9	0.09	0.09
	Protein	3.4mm	SNV+SG1D(15)	13	0.13	0.13
	Lactose	1.5mm	SNV+SG2D(19)	18	0.11	0.12
miniS-milk	Fat	T20	SNV+SG1D(21)	6	0.22	0.22
	Protein	T25	SNV	10	0.09	0.11
	Lactose	T25	SNV+SG1D(19)	11	0.09	0.10
miniS-sugarcane	Crude protein	NIR2.2k	SNV+SG2D(201)	5	0.62	0.63
	Total sugars	NIR2.2k	SNV	13	1.71	2.58

LV = number of latent variables; RMSECV = Root-mean-square error of cross-validation; RMSEP = Root-mean-square error of prediction; SNV = standard normal variates; SG1D(x) and SG2D(x) = first and second Savitzky-Golay derivatives, calculated using a second-order polynomial and a window size of x nm; mean centering is applied in all cases as the last pre-processing step.

In the SRS-milk dataset, each illumination-to-detection distance served as an individual block for single-block analysis. Best predictions were achieved with RMSEP values of 0.09%, 0.13% and 0.12% for respectively fat, protein and lactose. The efficacy of these predictions varied by distance. Optimal performances were obtained between 1.6 and 3.8 mm for fat and between 2.3 and 4 mm for protein, with peak accuracies at 2.3 mm and 3.4 mm, respectively. For lactose, optimal performance occurred between 1.1 and 1.8 mm, peaking at 1.5 mm. The distance ranges resulting in optimal performances had squared residuals that were not significantly higher than those for the distance resulting in peak accuracy with the lowest RMSEP for the respective response variable of the dataset.

In the miniS-milk dataset, for the prediction of milk fat, no significant difference was found between the T20 and the T25 spectrometers. However, the T20 model was simpler with six latent variables and a lower RMSEP (0.22%), compared to the T25 model (0.24%) with seven latent variables. For protein and lactose,

the T25 model outperformed the other single-block models, achieving respective RMSEP values of 0.11 and 0.10%.

Finally, for the miniS-sugarcane dataset, both the NIR2.2k and NIR1.7k spectrometers yield comparable capabilities in predicting CP. However, the NIR2.2k model is favored due to its marginally superior performance (RMSEP = 0.63%) and simpler model with five latent variables, as opposed to the NIR1.7k model (0.68%) with 15 latent variables. Additionally, the NIR2.2k model also provides the best TS prediction, with an RMSEP of 2.58%, outperforming all other spectrometers. It is hypothesized that the enhanced prediction performance of the NIR2.2k for CP and TS could be attributed to its spectral range (1150 to 2150 nm). This wavelength region contains overtones of the C–H, C–N, and N–H bonds between 1600 and 1700 nm [49], as well as specifically the N–H bonds at 2055 nm [50], both of which are related to proteins. Furthermore, the presence of O–H bonds from crystalline sucrose around 1441 nm contributes to the prediction of the sugar content [51].

3.3 Evaluation of the PROSAC prediction models

Figures 4, 5, and 6 present the PROSAC analysis results for the three distinct datasets. Specifically, Figure 4 illustrates RMSECV and RMSEP variations for fat, protein, and lactose in the SRS-milk dataset, along with the selection order of the different blocks used to build the PROSAC models. For this dataset, the algorithm manages 180 blocks, generated from six pre-processing methods and 30 illumination-to-detection distances, with a maximum of 50 blocks chosen for model building, indicating inevitable block repetition.

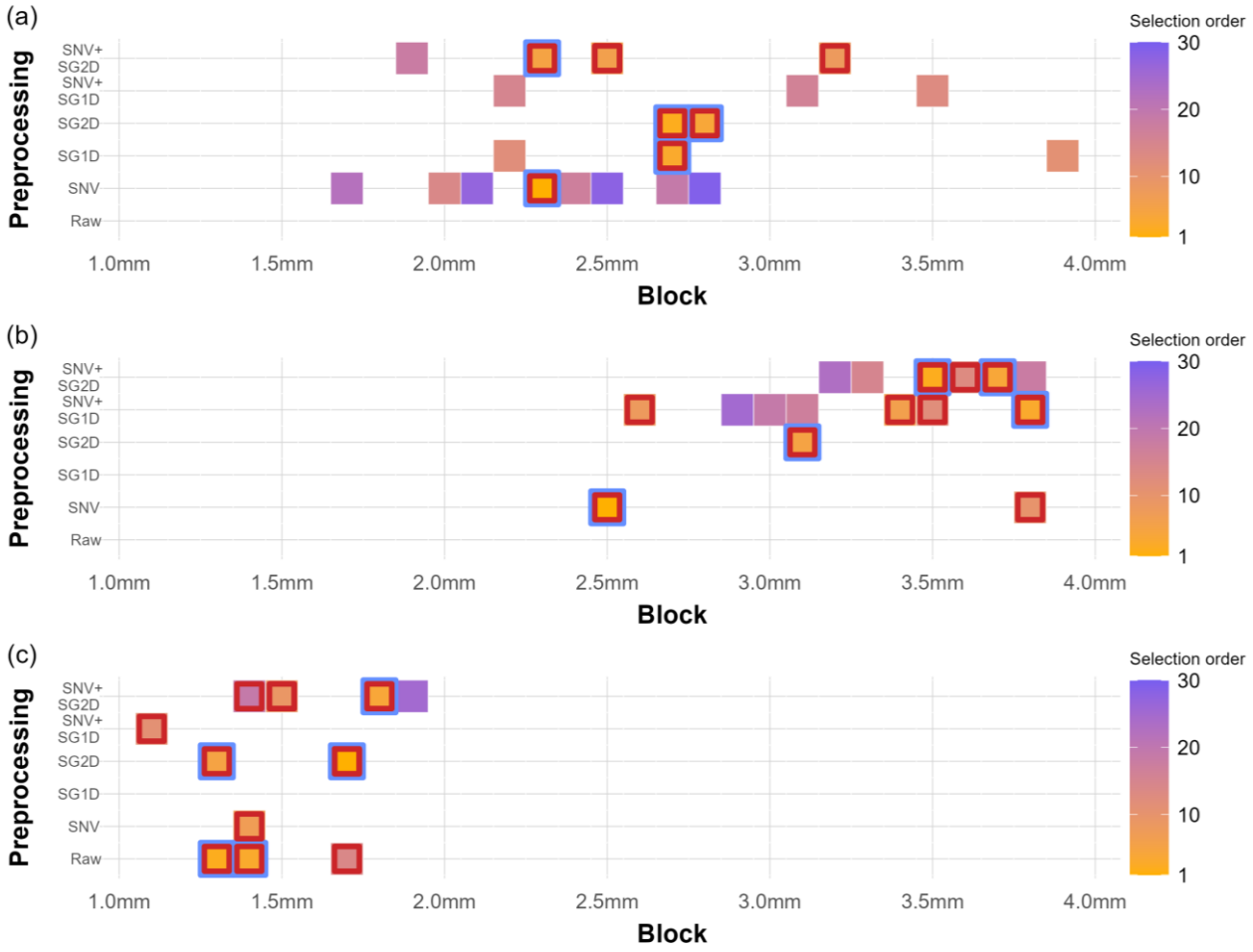


Figure 4. *PROSAC performance on the SRS-milk dataset for fat (a), protein (b) and lactose (c). The order in which the different blocks were selected is indicated by a yellow-to-purple gradient. Blocks chosen multiple times retain the color of their initial selection. Blue borders indicate the first five unique blocks selected by SO-PLS, while red highlights denote unique blocks chosen by PROSAC. SNV = Standard Normal Variates; SG2D = Savitzky-Golay second-order derivative; SG1D = Savitzky-Golay first-order derivative.*

In the case of fat prediction, a 10-block ensemble is selected by PROSAC, making use of seven unique blocks corresponding to five different distances of which some blocks are repeated. These seven unique blocks correspond to four of the six pre-processing methods, thereby confirming the ensemble model

construction capability of the PROSAC algorithm. The selected blocks (2.3 to 3.2 mm) fall within the statistically equivalent range of 1.6 to 3.8 mm in the single-block analysis. Notably, the 2.5 mm block pre-processed by an SNV followed by a SG2D is used four times in this 10-block ensemble, highlighting its role in providing complementary information despite not being among the top five blocks explaining the most variance. For the SO-PLS model, five unique blocks are selected: two each at 2.3 mm and 2.7 mm, and one at 2.8 mm, making use of selections of SNV, SG1D and SG2D, or a combination of SNV and the SG2D.

The PROSAC model for protein prediction comprises 14 blocks, using 10 unique blocks, predominantly utilizing SNV in combination with SG1D and SG2D. The first block selection corresponds to the 2.5 mm distance with SNV pre-processing, while the remaining significant blocks are confined to a 3.1 to 3.8 mm range. This closely aligns with the distance range resulting in optimal performances in the single-block analysis (2.3 to 4 mm). For SO-PLS implementation, the initial block at 2.5 mm with SNV pre-processing is followed by selections at 3.5, 3.8, 3.7, and 3.1 mm, employing SG2D and combinations of SNV with SG1D and SG2D.

When predicting lactose, the PROSAC model employs 20 blocks, using 10 unique blocks and incorporating five of the six available pre-processing methods. All selected blocks come from the 1.1 to 1.9 mm range, which largely overlaps with the distance range providing optimal performances in the single-block analysis. The 1.7 mm with a SG2D block, which was chosen as the first component of the model, was selected up to 17 times by the algorithm, emphasizing its significant role in predicting the lactose content in the milk samples. For the SO-PLS part of the methodology, the initial input includes this 1.7 mm block, followed by selections at 1.3 and 1.4 mm with no pre-processing, and a 1.8 mm block employing a combination of SNV and SG2D derivative.

A comparative analysis against the best single-block models reveals that the PROSAC implementation achieves similar performance metrics for fat (RMSEP = 0.09%, N = 10, being N the number of blocks

employed by PROSAC), protein (RMSEP = 0.13%, N = 14), and lactose (RMSEP = 0.12%, N = 20) in the SRS-milk dataset. The primary advantage of applying PROSAC in this context can be considered as simply saving time by developing a single PROSAC model compared to generating a multitude of single-block PLS models to find the best pre-processing method. More importantly, this approach allowed to select and sequence the blocks for subsequent implementation in the SO-PLS model. Particularly, for each of the three response variables, the distances selected by PROSAC align with the previously identified regions of interest for the accurate determination of fat, protein, and lactose in raw milk. These regions are 1.6 to 3.8 mm for fat, 2.3 to 4 mm for protein, and 1.1 to 1.8 mm for lactose [34].

The PROSAC results for the miniS-milk dataset are illustrated in Figures 5.a, 5.b, and 5.c where the block selection sequence for fat, protein, and lactose determination is outlined. With six pre-processing types and four spectrometers, PROSAC manages 24 individual blocks, selecting a maximum of 50 for model construction.

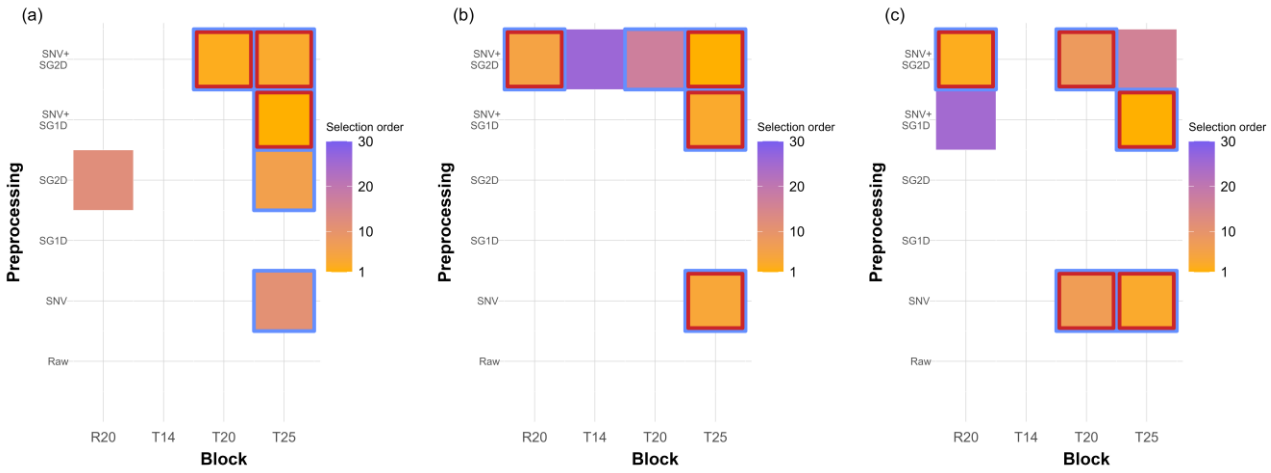


Figure 5. *PROSAC performance on the miniS-milk dataset for fat (a), protein (b) and lactose (c). The order in which the different blocks were selected is indicated by a yellow-to-purple gradient. Blocks chosen multiple times retain the color of their initial selection. Blue squares indicate the first five unique blocks selected by SO-PLS, while red highlights denote unique blocks chosen by*

PROSAC. T14 = NIRONE 1.4 in transmittance mode; R20 = NIRONE 2.0 in reflectance mode; T20 = NIRONE 2.0 in transmittance mode; T25 = NIRONE 2.5 in transmittance mode; SNV = Standard Normal Variates; SG2D = Savitzky-Golay second-order derivative; SG1D = Savitzky-Golay first-order derivative.

For fat content prediction in the miniS-milk dataset, the minimum RMSECV of the PROSAC model is achieved using only three blocks, without block repetition. These blocks combine T25 and T20 spectrometers and employ pre-processing methods such as SNV, SG1D and SG2D. This coincides with the findings in the single-block approach, where these two spectrometers had no significant difference in performance. Up to five non-repeated blocks are chosen for SO-PLS implementation, involving other pre-processings of T25 and exceeding the optimal three, under the hypothesis that SO-PLS will discard those blocks that do not contribute to an RMSECV reduction.

For protein estimation, the PROSAC model assembled up to 13 blocks comprising four unique blocks and mainly featuring repeated combinations of T25 and R20 blocks with diverse SNV and Savitzky-Golay derivative configurations. In the single-block approach, T25 gave the best performance, while R20 produced the second-best results (RMSECV = 0.14%, RMSEP = 0.14%, with nine latent variables), confirming its importance for the prediction of the protein content. Apart from the four unique blocks selected by PROSAC, the input for miniS-milk protein in the SO-PLS model also includes the T20 block with SNV pre-processing, which was selected as the 17th block by PROSAC.

Lactose prediction via PROSAC employs an 11-block ensemble, with five unique blocks incorporating T25, R20, and T20 blocks, mainly featuring SNV applications and in combination with Savitzky-Golay derivatives. This selection diverges slightly from the single-block analysis, where R20 resulted in a worse lactose prediction compared to transmittance, likely because of a lower interaction between reflected NIR light

and lactose molecules in the milk serum [41,45]. However, the wavelength range of the R20 and T20 spectrometers (1550 to 1950 nm) overlaps with the absorption bands of lactose linked to the O-H and C-H stretching vibrations [52], which may contribute to the prediction of milk lactose. Selection for SO-PLS involves the use of all these spectrometers and the previously mentioned pre-processings.

A comparative evaluation against the outcome of the single-block models indicates that the PROSAC performance metrics demonstrate a notable improvement, with the RMSEP for fat reduced from 0.22% to 0.19% (N = 3) and for protein from 0.11% to 0.10% (N = 13) in the miniS-milk dataset. For lactose, PROSAC achieved a similar performance with an RMSEP of 0.1% (N = 11), matching the single-block model. It must be noted that the T14 spectrometer block was never part of the blocks selected by PROSAC for model building. This is probably because its wavelength range (1100 to 1400 nm) overlaps with the 2nd and 3rd overtone bands of the milk components, while the other detectors can measure the stronger 1st overtone and combination bands.

The block selection sequence for constructing the PROSAC model for CP and TS determination from the miniS-sugarcane dataset is outlined in Figures 6.a and 6.b, respectively. Given six pre-processing methods and six spectrometers, PROSAC handles 36 individual blocks, selecting up to 50 for model construction.

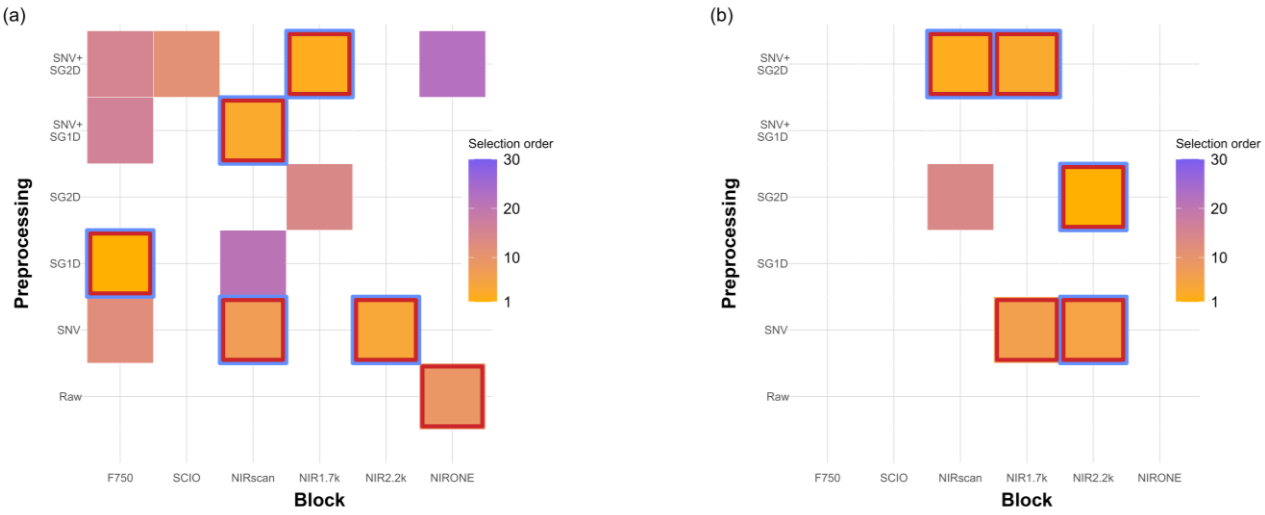


Figure 6. *PROSAC performance on the miniS-sugarcane dataset for crude protein (CP, a) and total sugar (TS, b). The order in which the different blocks were selected is indicated by a yellow-to-purple gradient. Blocks chosen multiple times retain the color of their initial selection. Blue squares indicate the first five unique blocks selected by SO-PLS, while red highlights denote unique blocks chosen by PROSAC. SNV = Standard Normal Variates; SG2D = Savitzky-Golay second-order derivative; SG1D = Savitzky-Golay first-order derivative.*

In the case of CP prediction, the PROSAC model incorporates 17 blocks, with six unique blocks combining all miniature spectrometers with all pre-processings except the singular application of a SG2D. In the single-block approach, the NIR2.2k and NIR1.7k spectrometers yielded the most accurate predictions, with RMSEP values of 0.63% and 0.64%, respectively. Nevertheless, the F750 spectrometer with a SG1D is chosen by PROSAC as the initial block, likely because its range (450 to 1140 nm) contains wavelengths pertinent to protein (1007 nm) and primary amines (1000 and 1020 nm) [53]. Additionally, the inclusion of NIR2.2k and NIR1.7k could have been anticipated from their superior performance in the single-block approach. The NIRscan is also amongst the primary selected blocks, with its spectral range (901 to 1701 nm) coinciding with that of the NIR1.7k spectrometer [53]. The NIRscan block pre-processed with SNV followed by a SG1D, and the NIRscan block with only SNV pre-processing, are selected as the third and fifth blocks in the PROSAC model. Although not primary selections, the SCIO (740 to 1070 nm) and NIRONE 2.2 (1750 to 2150 nm) spectrometers are also included by PROSAC, the former with a spectral region close to the F750 spectrometer and the latter being potentially relevant for protein-related N–H bonds at 2055 nm. The selection of five unique blocks for SO-PLS implementation contains data from the F750, NIR2.2k, NIRscan and NIR1.7k spectrometers.

For TS prediction, a seven-block ensemble with five unique blocks achieves the best performance featuring a mix of NIR2.2k, NIRscan, and NIR1.7k spectrometers with the application of SNV, SG2D and the combination of both pre-processings. The NIR2.2k spectrometer covers the C-O and H-O absorption bands around 2100 nm related to sugar content[53], while the spectral ranges of all the other selected spectrometers cover the region of the O-H bond from sugar as crystalline sucrose, around 1441 nm. These five unique blocks conform the input for the SO-PLS step.

Upon comparison with the best single-block models, the PROSAC implementation achieves similar performance metrics for both CP (RMSEP = 0.65%, N = 17) and TS (RMSEP = 2.58%, N = 7) in the MB-sugar dataset. Table 3 contains a summary of the prediction performance of PROSAC for each response variable and the blocks selected for SO-PLS implementation.

Table 3. Prediction performance statistics for the PROSAC prediction models for each of the response variables, for all datasets and selection of blocks for SO-PLS input.

Dataset	Response variable	*N*	RMSECV (% wt/wt)	RMSEP (% wt/wt)	Five unique blocks selected for SO-PLS input°
SRS-milk	Fat	10	0.09	0.09	B-2.3 mm, D-2.7 mm, C-2.7 mm, D-2.8 mm, F-2.3 mm
	Protein	14	0.10	0.13	B-2.5 mm, F-3.5 mm, E-3.8 mm, F-3.7 mm, D-3.1 mm
	Lactose	20	0.12	0.12	D-1.7 mm, A-1.3 mm, A-1.4 mm, F-1.8 mm, D-1.3 mm
miniS-milk	Fat	3	0.21	0.19	E-T25, F-T20, F-T25, D-T25, B-T25
	Protein	13	0.10	0.10	F-T25, E-T25, B-T25, F-R20, F-T20
	Lactose	11	0.10	0.10	E-T25, F-R20, B-T25,E-T25, B-T20
miniS-sugarcane	Crude protein	17	0.65	0.65	C-F750, F-NIR1.7k, E-NIRscan, B-NIR2.2k, B-NIRscan
	Total sugars	7	2.57	2.58	D-NIR2.2k, F-NIRscan, F-NIR1.7k, B-NIR2.2k, B-NIR1.7k

N = number of blocks used to build the model; RMSECV = Root-mean-square error of cross-validation; RMSEP = Root-mean-square error of prediction.

° For the input in SO-PLS letters A to F correspond to the following pre-processings: (A) Raw, (B) Standard Normal Variates, (C, D) Savitzky-Golay first or second order derivative respectively, and (E, F) Standard Normal Variates followed by Savitzky-Golay first or second order derivative respectively.

3.4 Evaluation of the PROSAC-SO-PLS prediction models

The performance metrics of the SO-PLS models starting from the blocks selected by PROSAC are detailed in Table 4. The last column of the table presents the blocks that were retained by the SO-PLS model from the first five unique blocks selected by PROSAC. Table 5 displays a summary of the RMSEP for the different approaches, comparing their performance with a paired test.

Table 4. Prediction performance statistics of the SO-PLS prediction models for the different response variables in the studied datasets.

Dataset	Response variable	LV	RMSECV (% wt/wt)	RMSEP (% wt/wt)	Selection by SO-PLS output°
SRS-milk	Fat	[5, 4]	0.08	0.09	[F-2.3 mm, B-2.3 mm]
	Protein	[12, 4]	0.09	0.10	[B-2.5 mm, F-3.5 mm]
	Lactose	[10, 2]	0.09	0.10	[A-1.3 mm, A-1.4 mm]
miniS-milk	Fat	[1, 1, 1]	0.21	0.19	[E-T25, F-T20, F-T25]
	Protein	[9, 4]	0.08	0.09	[B-T25, F-R20]
	Lactose	[14, 7]	0.08	0.08	[B-T20, E-T25]
miniS-sugarcane	Crude protein	[13, 3]	0.35	0.57	[F-NIR1.7k, C-F750]
	Total sugars	[16, 16, 6]	1.64	2.48	[D-NIR2.2k, F-NIR1.7k, B-NIR2.2k]

LV = number of latent variables used by the different blocks that comprise the SO-PLS model; RMSECV = Root-mean-square error of cross-validation; RMSEP = Root-mean-square error of prediction.

° For the selection by SO-PLS letters A to F correspond to the following pre-processings: (A) Raw, (B) Standard Normal Variates, (C, D) Savitzky-Golay first or second order respectively, and (E, F) Standard Normal Variates followed by Savitzky-Golay first or second order derivative, respectively.

Table 5. Prediction performance statistics of the different prediction models for the different response variables in the studied datasets.

Dataset	Response variable	RMSEP$^{\Delta}$ Single Block (% wt/wt)	RMSEP$^{\Delta}$ PROSAC (% wt/wt)	RMSEP$^{\Delta}$ PROSAC-SOPLS (% wt/wt)
SRS-milk	Fat	0.09	0.09	0.09
	Protein	0.13^{b}	0.13^{b}	0.10^{a}
	Lactose	0.12^{b}	0.12^{b}	0.10^{a}
miniS-milk	Fat	0.22^{b}	0.19^{a}	0.19^{a}
	Protein	0.11^{b}	0.10^{a,b}	0.09^{a}
	Lactose	0.10^{b}	0.10^{b}	0.08^{a}
miniS-sugarcane	Crude protein	0.63^{b}	0.65^{b}	0.57^{a}
	Total sugar	2.58^{b}	2.58^{b}	2.48^{a}

RMSEP = Root-mean-square error of prediction.

$^{\Delta}$Within each column, differing superscripts on RMSEP-values indicate significant differences (α = 0.05) between models per Tukey's HSD test; a lower alphabetical letter indicates a superior model.

The PROSAC methodology significantly reduced the RMSEP values for fat and protein prediction in the miniS-milk database, while the values for lactose prediction in this dataset and the parameters in the other datasets were comparable to those obtained with the single-block PLS approach, without statistical differences. As previously discussed, in this context PROSAC primarily benefits the modeler by streamlining the pre-processing selection and block selection for the SO-PLS step, thus significantly reducing the time required for model development compared to the iterative construction and evaluation of multiple single-block PLS models.

When applying the PROSAC-SO-PLS on the SRS-milk dataset, the prediction accuracy of the protein and lactose content improved significantly compared to both the single-block and PROSAC approach, with RMSEP values decreasing from 0.13% to 0.10% for protein and from 0.12% to 0.10% for lactose. However, the fat prediction did not show improvement, maintaining an RMSEP of 0.09% for all approaches. For fat,

the model employs the same distance as the best single-block approach, using two 2.3 mm blocks pre-processed with SNV and a SG2D. These blocks use four and five latent variables, respectively, matching the total used in the single-block approach. The PROSAC-SO-PLS protein model combines two distances (2.5 and 3.5 mm) with pre-processings similar to those used for fat. Notably, the 2.5 mm distance, although not selected as best in the single-block model, contributes significantly as indicated by its higher number of latent variables. The PROSAC-SO-PLS lactose prediction employs two adjacent distances (1.3 and 1.4 mm) without pre-processing.

For the miniS-milk dataset, the prediction of the response variables with the PROSAC-SO-PLS models shows marked improvement compared to the single-block approaches. For fat prediction, the RMSEP improved from 0.22% in the single-block approach to 0.19% in both PROSAC and PROSAC-SO-PLS. For this milk component, the first three blocks chosen by PROSAC are also retained by SO-PLS, each utilizing a single latent variable and achieving identical performance metrics as PROSAC, which is also significantly better than the outcome of the single-block approach. For protein, SO-PLS selects only the third and fourth blocks from the PROSAC output, slightly improving the RMSEP from 0.11% in the single-block to 0.10% in PROSAC, and further to 0.09% in PROSAC-SO-PLS. In lactose prediction, where transmittance-based miniature spectrometers excel, the PROSAC-SO-PLS model outperformed both the single-block approach (RMSEP reduced from 0.10% to 0.08%) and PROSAC (RMSEP at 0.10%), despite increased model complexity.

In the miniS-sugarcane dataset, both CP and TS predictions with PROSAC-SO-PLS showed significant improvements compared to PROSAC and the best single-block models. For CP, the RMSEP improved from 0.63% in the single-block and 0.65% in PROSAC to 0.57% in PROSAC-SO-PLS. The F750 and NIR1.7k miniature spectrometers effectively capture wavelengths related to protein and primary amines. The inclusion of protein-related N–H bonds at 2055 nm does appear to add value to the models as SO-PLS did not retain the NIR2.2k blocks for CP prediction. This could be attributed to the high moisture content in

the sugarcane samples, which likely results in a low signal-to-noise ratio at longer wavelengths, thereby reducing the effectiveness of these spectral features in the model. For TS prediction, a combination of NIR1.7k and NIR2.2k yields superior results, with the RMSEP decreasing from 2.58% in both the single-block and PROSAC to 2.48% in PROSAC-SO-PLS. The NIR1.7k covers the first overtone of O-H stretching around 1441 nm, while the NIR2.2k captures data around 2100 nm, particularly focusing on the combination bands arising from C–O stretching and O–H bending, indicators of sugar content.

In the SO-PLS model-building strategy, if any of the initial five input blocks had zero latent variables, subsequent iterations explored the inclusion of additional PROSAC blocks. Despite these adjustments, no statistically significant improvement in RMSECV ($\alpha = 0.05$) was observed, rendering the strategy unnecessary. Importantly, all SO-PLS models used no more than three blocks and never involved more than two different devices or illumination-to-detection distances. This suggests that the selection from the initial five PROSAC blocks was sufficient for compositional predictions across all datasets examined. Additionally, while the models demonstrated good agreement with our calibration and test split, future assessments should incorporate completely independent external test sets, separate from a common dataset split. This approach will provide a more robust evaluation of the performance of the procedure, ensuring its validity and generalizability across diverse sample sets.

The computational cost varied across different methods, depending on the dataset. The SRS-milk dataset presented the highest complexity, requiring the construction of 30 single-block PLSR models for each of the 30 illumination-to-detection distances, considering up to 20 latent variables. For the single-block approach, testing all pre-processing combinations across these distances took approximately 200 seconds in Matlab 2021a on the previously defined hardware. In contrast, the PROSAC algorithm processed 180 blocks, derived from applying six pre-processing methods to the 30 distances. It utilized a maximum of 50 blocks for model selection and required about 210 seconds on the same computer to optimize, develop, and test a final PROSAC model. However, the SO-PLS step, considering 20 latent variables, is the most

computationally demanding. The time to construct and apply a prediction model grows exponentially with the increase in input blocks. For instance, restricting the SO-PLS input to three blocks took around 300 seconds with the same hardware, while employing five blocks extended to about 26 hours. The use of six blocks theoretically demands approximately 35 days. This underscores the critical role of PROSAC in efficiently selecting the most relevant blocks to significantly reduce the SO-PLS runtime. It also highlights the necessity of minimizing the number of blocks fed into the SO-PLS model to ensure manageable computational times.

4 Conclusion

This study introduced the PROSAC-SO-PLS methodology, which integrates PROSAC, a selection and ordering method for pre-processed data blocks, with SO-PLS, which allows for a more targeted ensemble model construction. By limiting the selection to a maximum of five most explanatory blocks, the approach can build accurate prediction models with reduced complexity. This can be vital when trying to minimize the number of physical components in sensor system development. Our analysis of three NIR datasets, encompassing eight different response variables, demonstrated that the PROSAC-SO-PLS method surpassed traditional approaches in seven out of eight variables, achieving a reduction in RMSEP ranging from 5 to 25%, with a maximum combination of three blocks for all cases. This improvement was attributed to the ensemble use of differently pre-processed NIR data and the targeted, ordered implementation in SO-PLS.

In summary, the PROSAC-SO-PLS methodology offers a robust and efficient approach to multiblock modeling in spectroscopy, transcending the specific context of NIR spectroscopy. This method addresses the limitations of existing techniques across various spectroscopic domains. By integrating the strengths of PROSAC and SO-PLS, this study paves the way for more targeted and computationally efficient predictive models. The methodology not only excels in performance but also significantly reduces the complexity of the model, making it a viable option for both experimental and commercial applications of multi-block modeling.

5 Acknowledgments

José A. Diaz Olivares was funded by the Research Foundation Flanders (FWO, Belgium) through the PhD fellowship strategic basic research No. 1S76322N and the travel grant No. V418621N. Ines Adriaens is funded via the special research fund (BOF) of Ghent University. Additionally, this research has been

financially supported by KU Leuven internal funding (C3 project C3/19/037). No conflicts of interest have been reported by the authors.

6 References

[1] M. Cocchi, Introduction: Ways and Means to Deal With Data From Multiple Sources, Data Handling in Science and Technology 31 (2019) 1–26. https://doi.org/10.1016/B978-0-444-63984-4.00001-6.

[2] S. Mayr, K.B. Beć, J. Grabska, E. Schneckenreiter, C.W. Huck, Near-infrared spectroscopy in quality control of Piper nigrum: A comparison of performance of benchtop and handheld spectrometers, Talanta 223 (2021) 121809. https://doi.org/10.1016/j.talanta.2020.121809.

[3] J. Riu, G. Gorla, D. Chakif, R. Boqué, B. Giussani, Rapid Analysis of Milk Using Low-Cost Pocket-Size NIR Spectrometers and Multivariate Analysis, Foods 2020 9 (2020) 1090. https://doi.org/10.3390/foods9081090.

[4] G. Stocco, C. Cipolat-Gotet, A. Ferragina, P. Berzaghi, G. Bittante, Accuracy and biases in predicting the chemical and physical traits of many types of cheeses using different visible and near-infrared spectroscopic techniques and spectrum intervals, Journal of Dairy Science 102 (2019) 9622–9638. https://doi.org/10.3168/jds.2019-16770.

[5] H. Yaman, D.P. Aykas, R. Jiménez-Flores, L.E. Rodriguez-Saona, Monitoring the ripening attributes of Turkish white cheese using miniaturized vibrational spectrometers, Journal of Dairy Science 105 (2022) 40–55. https://doi.org/10.3168/jds.2021-20313.

[6] D. Lahat, T. Adali, C. Jutten, Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects, Proceedings of the IEEE 103 (2015) 1449–1477. https://doi.org/10.1109/jproc.2015.2460697.

[7] I. Måge, B.H. Mevik, T. Næs, Regression models with process variables and parallel blocks of raw material measurements, Journal of Chemometrics 22 (2008) 443–456. https://doi.org/10.1002/cem.1169.

[8] R.A. Crocombe, Portable Spectroscopy, Applied Spectroscopy 72 (2018) 1701–1751. https://doi.org/10.1177/0003702818809719

[9] J.S. Shenk, J.J. Workman, M.O. Westerhaus, Application of NIR spectroscopy to agricultural products, Handbook of Near-Infrared Analysis, Third Edition (2007) 347–386. https://doi.org/10.1201/9781420007374-24

[10] K.B. Beć, J. Grabska, C.W. Huck, Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives, Foods 2022 11 (2022) 1465. https://doi.org/10.3390/foods11101465.

[11] L. Zhou, C. Zhang, Z. Qiu, Y. He, Information fusion of emerging non-destructive analytical techniques for food quality authentication: A survey, Trends in Analytical Chemistry 127 (2020) 115901. https://doi.org/10.1016/j.trac.2020.115901.

[12] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Analytica Chimica Acta 891 (2015) 1–14. https://doi.org/10.1016/j.aca.2015.04.042.

[13] E. Hayes, D. Greene, C. O'Donnell, N. O'Shea, M.A. Fenelon, Spectroscopic technologies and data fusion: Applications for the dairy industry, Frontiers in Nutrition 9 (2023) 1074688. https://doi.org/10.3389/fnut.2022.1074688.

[14] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

[15] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[16] A.K. Smilde, I. Måge, T. Næs, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, Journal of Chemometrics 31 (2017) e2900. https://doi.org/10.1002/cem.2900.

[17] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, Trends in Analytical Chemistry 137 (2021) 116206. https://doi.org/10.1016/j.trac.2021.116206.

[18] A.K. Smilde, I. Van Mechelen, A Framework for Low-Level Data Fusion, Data Handling in Science and Technology 31 (2019) 27–50. https://doi.org/10.1016/B978-0-444-63984-4.00002-8.

[19] A. Smolinska, J. Engel, E. Szymanska, L. Buydens, L. Blanchet, General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences, Data Handling in Science and Technology 31 (2019) 51–79. https://doi.org/10.1016/B978-0-444-63984-4.00003-X.

[20] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, Journal of Chemometrics 25 (2011) 28–40. https://doi.org/10.1002/cem.1357.

[21] A. Biancolillo, T. Næs, The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions, Data Handling in Science and Technology 31 (2019) 157–177. https://doi.org/10.1016/B978-0-444-63984-4.00006-5.

[22] A.K. Smilde, T. Næs, K.H. Liland, Multiblock data fusion in statistics and machine learning: Applications in the natural and life sciences, John Wiley & Sons, (2022).

[23] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, Chemometrics and Intelligent Laboratory Systems 124 (2013) 32–42. https://doi.org/10.1016/j.chemolab.2013.03.006.

[24] T. Næs, R. Romano, O. Tomic, I. Måge, A. Smilde, K.H. Liland, Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects, Journal of Chemometrics 35 (2021) e3243. https://doi.org/10.1002/cem.3243.

[25] M. P. Campos, R. Sousa, M. S. Reis, Establishing the optimal blocks' order in SO-PLS: Stepwise SO-PLS and alternative formulations, Journal of Chemometrics 32 (2018) e3032. https://doi.org/10.1002/cem.3032.

[26] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trends in Analytical Chemistry 132 (2020) 116045. https://doi.org/10.1016/j.trac.2020.116045.

[27] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends in Analytical Chemistry 28 (2009) 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

[28] J.M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems 199 (2020) 103975. https://doi.org/10.1016/j.chemolab.2020.103975.

[29] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, Journal of Chemometrics 30 (2016) 651–662. https://doi.org/10.1002/cem.2824.

[30] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling, Chemometrics and Intelligent Laboratory Systems 222 (2022) 104497. https://doi.org/10.1016/j.chemolab.2022.104497.

[31] P. Mishra, Bypassing NIR pre-processing optimization with multiblock pre-processing ensemble approaches, NIR news 33 (2022) 5–8. https://doi.org/10.1177/09603360221139227.

[32] H.R. Cederkvist, A.H. Aastveit, T. Næs, A comparison of methods for testing differences in predictive ability, Journal of Chemometrics 19 (2005) 500–509. https://doi.org/10.1002/cem.956.

[33] A. Biancolillo, Method development in the area of multi-block analysis focused on food analysis : PhD thesis, (2016).

[34] J.A. Diaz-Olivares, M.J. Gote, W. Saeys, I. Adriaens, B. Aernouts, K.U. Leuven, Near-infrared spatially-resolved spectroscopy for milk quality analysis, (2023). https://doi.org/10.26434/CHEMRXIV-2023-KSWCT.

[35] ISO, Milk and Liquid Milk Products–Guidelines for the Application of Mid-Infrared Spectrometry, ISO Norm 9622: 2013/IDF 141: 2013 (2013).

[36] S. Uusitalo, J. A. Diaz-Olivares, J. Sumen, E. Hietala, I. Adriaens, W. Saeys, M. Utriainen, L. Frondelius, M. Pastell, B. Aernouts, Evaluation of MEMS NIR Spectrometers for On-Farm

Analysis of Raw Milk Composition, Foods 10 (2021) 2686. https://doi.org/10.3390/foods10112686.

[37] A. Zgouz, D. Héran, B. Barthès, D. Bastianelli, L. Bonnal, V. Baeten, S. Lurol, M. Bonin, J.M. Roger, R. Bendoula, G. Chaix, Dataset of visible-near infrared handheld and micro-spectrometers – comparison of the prediction accuracy of sugarcane properties, Data in Brief 31 (2020) 106013. https://doi.org/10.1016/j.dib.2020.106013.

[38] European Commission, Commission Regulation (EC) No 152/2009 of 27 January 2009 laying down the methods of sampling and analysis for the official control of feed, Official Journal of the European Union 54 (2009) 2–54.

[39] H.C.S. De Whalley, ICUMSA methods of sugar analysis: official and tentative methods recommended by the International Commission for Uniform Methods of sugar analysis (ICUMSA), Elsevier (2013). https://doi.org/10.1016/C2013-0-12079-X

[40] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics 19 (1977) 415–428. https://doi.org/10.1080/00401706.1977.10489581.

[41] J.A. Diaz-Olivares, I. Adriaens, E. Stevens, W. Saeys, B. Aernouts, Online milk composition analysis with an on-farm near-infrared sensor, Computers and Electronics in Agriculture 178 (2020) 105734. https://doi.org/10.1016/j.compag.2020.105734.

[42] M. Ryckewaert, G. Chaix, D. Héran, A. Zgouz, R. Bendoula, Evaluation of a combination of NIR micro-spectrometers to predict chemical properties of sugarcane forage using a multi-block approach, Biosystems Engineering 217 (2022) 18–25. https://doi.org/10.1016/j.biosystemseng.2022.02.019.

[43] W. Saeys, K. Beullens, J. Lammertyn, H. Ramon, T. Naes, Increasing Robustness against Changes in the Interferent Structure by Incorporating Prior Information in the Augmented Classical Least-Squares Framework, Analytical Chemistry 80 (2008) 4951–4959. https://doi.org/10.1021/ac800155N.

[44] J.A. Diaz-Olivares, A. van Nuenen, M.J. Gote, V.F. Díaz, W. Saeys, I. Adriaens, B. Aernouts, Near-infrared spectra dataset of milk composition in transmittance mode, Data in Brief 51 (2023) 109767. https://doi.org/10.1016/j.dib.2023.109767.

[45] B. Aernouts, E. Polshin, J. Lammertyn, W. Saeys, Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: Reflectance or transmittance?, Journal of Dairy Science 94 (2011) 5315–5329. https://doi.org/10.3168/jds.2011-4354.

[46] N. V. Hoang, A. Furtado, L. Donnan, E.C. Keeffe, F.C. Botha, R.J. Henry, High-Throughput Profiling of the Fiber and Sugar Composition of Sugarcane Biomass, Bioenergy Research 10 (2017) 400–416. https://doi.org/10.1007/S12155-016-9801-8.

[47] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR spectroscopy with applications in food and beverage analysis., Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. (1993). H. Longman Scientific and Technical, Harlow, Essex, UK. ISBN 0582099463.

[48] J.C. Tewari, K. Malik, In situ laboratory analysis of sucrose in sugarcane bagasse using attenuated total reflectance spectroscopy and chemometrics, International Journal of Food Science & Technology 42 (2007) 200–207. https://doi.org/10.1111/J.1365-2621.2006.01209.X.

[49] D.H. Clark, R.C. Lamb, Near Infrared Reflectance Spectroscopy: A Survey of Wavelength Selection To Determine Dry Matter Digestibility, Journal of Dairy Science 74 (1991) 2200–2205. https://doi.org/10.3168/JDS.S0022-0302(91)78393-8.

[50] D.L. Wetzel, Near-Infrared Reflectance Analysis: Sleeper Among Spectroscopic Techniques, Analytical Chemistry 55 (1983) 1165–1176. https://doi.org/10.1021/AC00262A001.

[51] S.E. Kays, F.E. Barton, W.R. Windham, D.S. Himmelsbach, Prediction of Total Dietary Fiber by Near-Infrared Reflectance Spectroscopy in Cereal Products Containing High Sugar and Crystalline Sugar, Journal of Agricultural and Food Chemistry 45 (1997) 3944–3951. https://doi.org/10.1021/JF9703260.

[52] O. Scheibelhofer, P. Wahl, B. Larchevêque, F. Chauchard, J. Khinast, Spatially Resolved Spectral Powder Analysis: Experiments and Modeling, Applied Spectroscopy 72 (2018) *521-534*. https://doi.org/10.1177/0003702817749839.

[53] J. Workman, A.W. Springsteen, Applied Spectroscopy: A Compact Reference for Practitioners, (1998). https://doi.org/10.1016/B978-0-12-764070-9.X5000-8.