Chemometrics and Intelligent Laboratory Systems 222 (2022) 104497

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

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

Puneet Mishra [a,*], Jean Michel Roger [b,c], Federico Marini [d], Alessandra Biancolillo [e], Douglas N. Rutledge [f,g]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France
[d] Department of Chemistry, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy
[e] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, 67100, Coppito, L'Aquila, Italy
[f] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005, Paris, France
[g] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

ARTICLE INFO

Keywords:
Multi-block modelling
Pre-processing
Spectroscopy
Data fusion

1. ABSTRACT

Ensemble pre-processing is emerging as a potential tool to avoid the tiring pre-processing selection and optimization task in near-infrared (NIR) spectral modelling. Furthermore, differently pre-processed data may carry complementary information, hence, ensemble pre-processing may represent the best suited modelling option to extract all the useful information from differently pre-processed data. Recently, multi-block techniques such as sequential (SPORT) and parallel (PORTO) orthogonalized partial least squares regression were proposed to extract complementary information present in differently pre-processed data. Although such multi-block techniques allowed efficient modelling of differently pre-processed data blocks, depending on the approach, challenges related to choosing block order, parameter tuning, block scaling and optimization time requirements still must be dealt with. To cope with such issues, the present study proposes the use of a recently developed faster, block order independent and scale independent, multi-block data modelling technique called response-oriented sequential alternation (ROSA) to process the multi-block data generated by differently pre-processing the same NIR data. This new method is called PROSAC, i.e., pre-processing ensembles with ROSA calibration. The potential of the approach is demonstrated on five real NIR spectral datasets. Furthermore, as baselines for comparison, partial least squares regression was done on individually pre-processed data sets, and using two multi-block pre-processing fusion approaches, i.e., SPORT and PORTO. The ensemble pre-processing with ROSA achieved either better performance compared to the baseline methods or achieved comparable performance without the need to worry about the pre-processing order, the scaling of data after pre-processing and optimization time requirements. PROSAC can be considered as a general tool for the ensemble pre-processing for NIR data modelling.

1. Introduction

Near-infrared (NIR) spectroscopy is a widely used non-destructive optical sensing technique often deployed for rapid and contact-less analysis of materials [1]. NIR is based on the interaction of infrared radiation with the materials and its consequent absorption, reflection, and transmission spectra are used to characterize the material properties [1,2]. Combined with chemometric processing, NIR spectroscopy can be used for both qualitative and quantitative analysis of samples [3,4]. Applications of NIR spectroscopy can be found ranging from agriculture [5] to high-end pharmaceutical manufacturing and process control [6]. Furthermore, NIR spectroscopy can be explored in either point or imaging spectroscopy mode, capturing the spatially resolved spectral properties of materials [1,7,8].

Although many applications of NIR spectroscopy can be found in the literature [1,9–13], a common struggle with regard to its proper

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.chemolab.2022.104497
Received 10 October 2021; Received in revised form 7 January 2022; Accepted 13 January 2022
Available online 15 January 2022

Table 1
A summary of different data sets used in this study. The data sets were partitioned in calibration and tests set using the Kennard-Stone algorithm.

Samples	Wavelength (nm)	Spectral variables	nCal/nTest	Reference content (%)	Calibration range (mean ± std)	Test range (mean ± std)
Mango	684–990	103	300/201	Moisture	85.36 ± 2.03	85.53 ± 1.85
Pear	720–997	85	330/221	Soluble solids	12.62 ± 1.34	12.53 ± 1.32
Apple	683–992	95	1210/808	Dry matter	15.49 ± 1.67	15.54 ± 1.48
Olive	720–999	94	349/234	Dry matter	28.47 ± 3.12	28.79 ± 3.23
Avocado	730–999	83	283/190	Dry matter	20.93 ± 2.71	20.66 ± 2.25

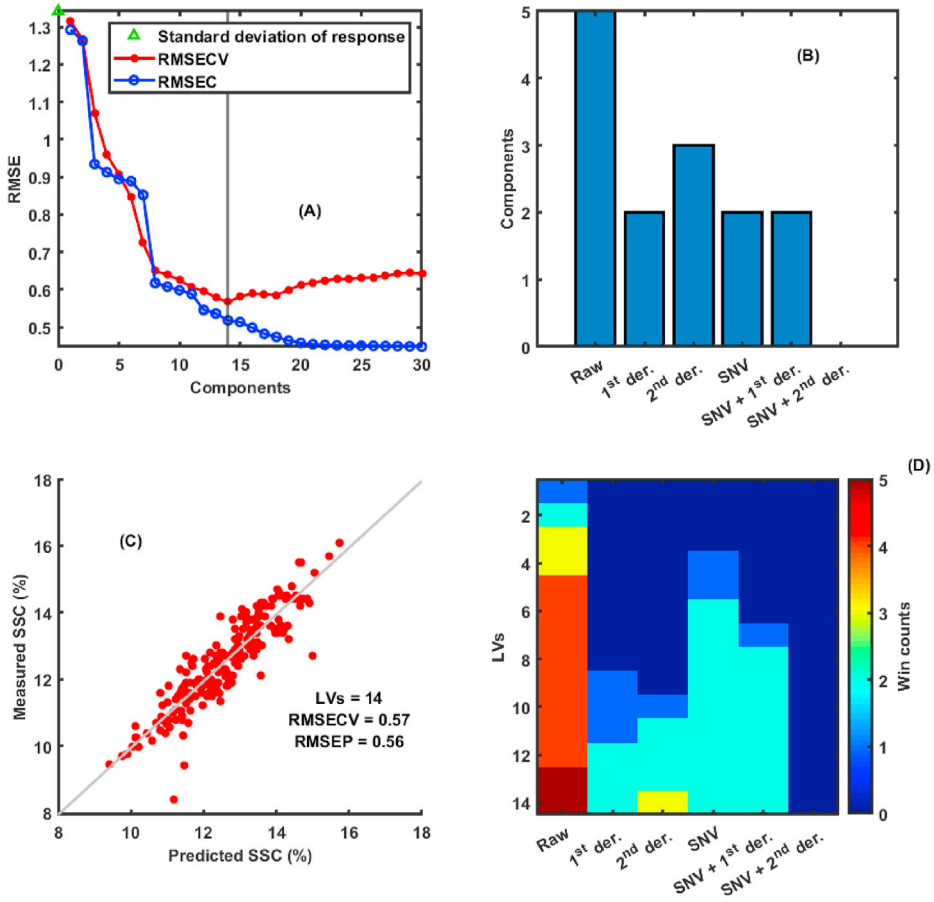

Fig. 1. Performance of PROSAC modelling on Pear data set. (A) Cross-validation plot for PROSAC, (B) LVs for each data block for PROSAC, (C) Prediction plot for PROSAC, and (D) Winning block order for LVs. In subplot (D), a change of colour along the vertical direction indicates that the latent variable is selected from that block. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
A summary of PLS models made on individually pre-processed data block for all data sets.

Data sets	Raw data (LVs/ RMSEP)	1st derivative (LVs/ RMSEP)	2nd derivative (LVs/ RMSEP)	SNV (LVs/ RMSEP)	SNV+1st derivative (LVs/ RMSEP)	SNV+2nd derivative (LVs/ RMSEP)
Pear	11/0.60	10/0.61	8/0.60	10/0.62	11/0.61	9/0.62
Avocado	10/1.32	10/1.29	10/1.26	10/1.26	10/1.27	10/1.22
Apple	9/0.73	9/0.71	8/0.72	8/0.74	10/0.73	8/0.71
Mango	11/0.61	12/0.57	11/0.55	10/0.75	9/0.76	8/0.72
Olive	8/1.53	6/1.54	7/1.56	7/1.37	4/1.50	6/1.40

implementation is the associated chemometric modelling required to calibrate the spectral sensors for the desired tasks, for example, for the prediction of a chemical constituent or for the classification of samples to predefined classes [3,14]. Classical latent variable modelling techniques such as partial least squares (PLS) [15,16] regression/discriminant analysis and advanced deep convolutional neural networks [17–19] are widely used. However, before performing the chemometric or deep learning modelling, the NIR data requires extra work in terms of removing the artefacts present in the signal, such as the additive and multiplicative effects caused by the light scattering [20–22]. Following its interaction with a material, the NIR light is subject to two major phenomena: absorption and scattering [14,23]. The absorption is due to the chemical components present in the samples while scattering is mainly related to the attenuation in the signal due to the interaction of light with the physical structure of the materials [1]. In general, when predicting chemical constituents, it is crucial that the model be solely

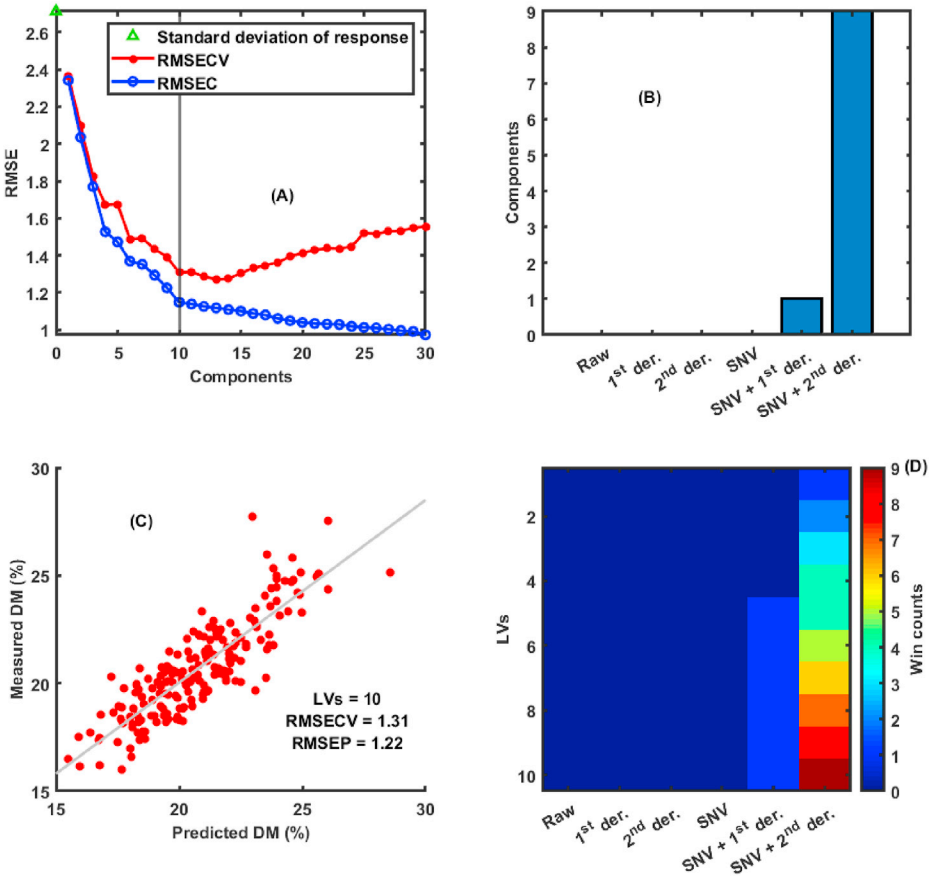

Fig. 2. Performance of PROSAC modelling on Avocado data set. (A) Cross-validation plot for PROSAC, (B) LVs for each data block for PROSAC, (C) Prediction plot for PROSAC, and (D) Winning block order for LVs. In subplot (D), a change of colour along the vertical direction indicates that the latent variable is selected from that block. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

based on the absorption characteristics in the NIR spectra and not on the scattering [20,21]. However, in cases where the scattering information may supply an additional correlation with the property of interest, it is recommended to use raw data [24,25]. In any case, before the NIR data modelling for qualitative and quantitative purposes, a wide exploration of different chemometric pre-processing techniques is needed to understand if it is required or not [21]. Some commonly used pre-processing techniques for NIR data modelling are normalisation techniques such as standard normal variate (SNV) [26], variable sorting for normalisation (VSN) [27], derivatives such as Savitzky-Golay (SavGol) [28], baseline correction [29], physically based models for scatter correction such as multiplicative scatter correction or its extended forms [30–32] and many more [20,21]. Furthermore, several pre-processing are also used in combinations such as normalisation techniques in combination with differentiation, where the normalisation attenuates light scatter effects, and the derivative reveals underlying peaks in the spectra [33].

The challenge of chemometric pre-processing choice and exploration is well recognized in the chemometrics community and several exhaustive approaches [34,35] and experimental design (DoE) based approaches [36,37] exist. These approaches, by exploring several pre-processing methods and their combinations, can find the ones best suited for a data set. The main drawback of such approaches is that they can easily become a time and computing resource-consuming task [20]. There are also faster exhaustive methods based on genetic PLS approaches [38]. The drawback of exhaustive approaches is that they aim to find a single pre-processing or a single combination of pre-processings, while different pre-processing and/or their combinations may carry complementary information which, if modelled in an ensemble strategy, could result in better model accuracies [20]. Several recent studies [6,25, 33,39–41] have shown the complementary nature of different pre-processings and highlighted the need for an ensemble pre-processing modelling.

In the domain of chemometrics, ensemble pre-processing approaches to NIR data modelling are emerging [20]. There are currently three main types of ensemble approaches: stacked regression, DoE based, and multi-block inspired. In the first one [42], several models are built based on different pre-processings and stacked for ensemble modelling; in the DoE based approach [43], a full-factorial design is explored for all combinations of pre-processings and the best performing models are then combined. The third approach is inspired by the family of multi-block data fusion approaches used in the chemometric community such as sequential and parallel orthogonalized partial least squares regression analysis [44,45], which treats the same NIR data after different pre-processings as a multi-block data set. Of the three approaches, the multi-block ensemble approaches are of particular interest as they retain all-important chemometric parameters such as regression coefficients, scores, and loading, which are of great interest for model interpretation and for understanding the background spectrochemistry of the models [25]. Currently, there are two main multi-block ensemble pre-processing approaches available, i.e., sequential pre-processing through orthogonalization (SPORT) [44] and parallel pre-processing through orthogonalization (PORTO) [45]. Both these approaches have been shown to outperform the selective pre-processing modelling approaches in several application areas [6,25,33,39–41].

Although both these approaches, i.e., SPORT [44] and PORTO [45] allow pre-processing ensembles modelling for NIR data, they do have their drawbacks from the operational and model optimization perspectives. The main drawback of these approaches is that they are highly dependent on

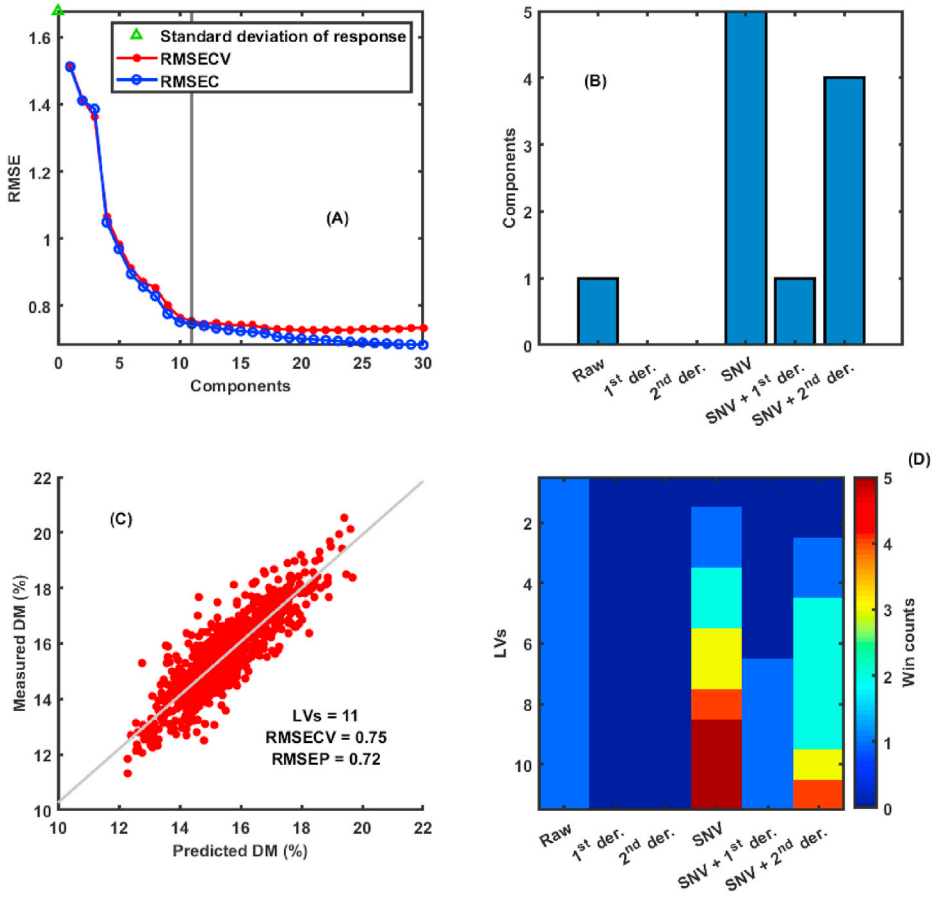

Fig. 3. Performance of PROSAC modelling on Apple data set. (A) Cross-validation plot for PROSAC, (B) LVs for each data block for PROSAC, (C) Prediction plot for PROSAC, and (D) Winning block order for LVs. In subplot (D), a change of colour along the vertical direction indicates that the latent variable is selected from that block. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the data block order, not necessarily in terms of prediction accuracy, but rather in terms of the values of the model coefficients and the resulting optimal complexity. For example, the SPORT approach, which is based on the concept of sequential and orthogonalized PLS (SO-PLS) regression, processes the data blocks sequentially; hence, if the order of the data blocks is changed the SPORT model can select different numbers of latent variables for each block and the results and interpretation will be modified. The PORTO approach is less dependent on data block order; however, it is at least partly affected, since the components which are common to some blocks only are extracted based on the order in which all the possible subgroups of blocks are explored. The second main drawback of SPORT and PORTO is the time required for model optimization; for example, the SPORT approach based on the SO-PLS requires exploration of all combinations of latent variables from different data blocks before selecting the optimal one. Such an exploration is workable when the number of blocks is low [46] but can become a very tedious task when the number of data blocks increases. Such a computation cost is a limitation for exploring a wide number of differently pre-processed blocks together. Hence, to deal with the two main drawbacks of the SPORT and PORTO approaches, this study proposes the implementation of the recently developed response-oriented sequential alternation (ROSA) method [46] for a faster order- and scale-independent ensemble modelling of several data blocks. ROSA [46] is a new multiblock extension of PLS regression which supplies all the relevant parameters such as regression coefficient, scores and loading.

This study aims to propose ROSA [46] as a novel tool for ensemble pre-processing modelling for NIR spectroscopy data. This new method is called PROSAC: pre-processing ensembles with ROSA calibration. To demonstrate this, the PROSAC approach was applied to five different data sets pre-processed with several different methods and the outcomes were compared to those of single-block PLS analyses on the individually pre-processed data blocks. Furthermore, a comparison with the SPORT [44] and PORTO [45] approaches is also presented in terms of prediction power and of the effect of the order of the data blocks. The time required by PROSAC to handle large number of blocks is also explored.

2. Materials and methods

2.1. Data sets

Five real NIR datasets were used to demonstrate the potential of PROSAC for ensemble pre-processing modelling as well for its comparison with the baseline techniques. The five datasets were related to the prediction of dry matter (DM %), soluble solids content (SSC %) and moisture content (MC %) in five different fresh fruits. All data sets were measured with Felix fresh fruit quality meter (Camas, WA, USA). The Felix fruit quality meter is a hand-held spectrometer that uses the interaction mode to measure the spectral signature of samples. The Felix spectrometer covers the spectral range of 310–1135 nm with a ~3 nm spectral sampling interval. The spectrometer has a Xenon Tungsten Lamp for illumination and a built-in white painted reference standard for estimating the reflectance. For four out of five data sets, the data has already been used in earlier publications by different authors around the world. For example, the Mango data set (cultivar Keitt and Kent) was the same as the validation set used in Ref. [47], the Pear data set (cultivar Conference) was the same as used in Ref. [48], the Apple data set (cultivar Cripps Pink, Fuji, Gala, Golden Delicious and Honeycrisp) was sourced from Ref. [49] and the Olive data set was sourced from Ref. [50].

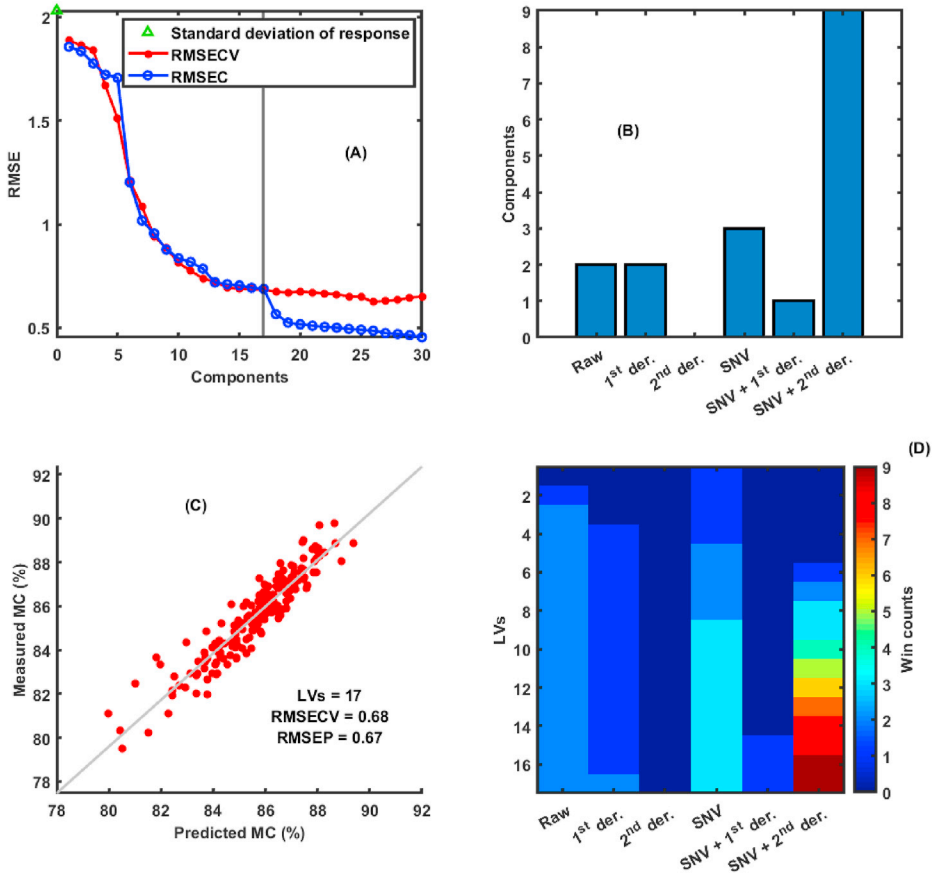

Fig. 4. Performance of PROSAC modelling on Mango data set. (A) Cross-validation plot for PROSAC, (B) LVs for each data block for PROSAC, (C) Prediction plot for PROSAC, and (D) Winning block order for LVs. In subplot (D), a change of colour along the vertical direction indicates that the latent variable is selected from that block. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The Avocado (cultivar Hass) data set was a new data set measured in this study and involved measurement of NIR spectra followed by dry matter measurements using the oven drying method of the avocado flesh. The fruit flesh was extracted from the same spot where the NIR measurements were performed. Avocado flesh was placed into clean aluminium cups and initial weights (g) were recorded using a balance (Mettler-Toledo GmbH, Giessen, Germany). Later, the samples were dried in a hot-air oven (FD 56, Binder GmbH, Tuttlingen, Germany) at 80°C for 60 h and the final weights (g) of the aluminium cups were recorded with an analytical balance (Mettler-Toledo GmbH, Giessen, Germany). A summary of all data sets is provided in Table 1. It can be noted that the wavelength ranges for different data sets varied slightly but all were in the 3rd overtones range of OH and CH and relevant for prediction of parameters such as dry matter, soluble solids content and moisture content. The reason for the slight variation in the wavelength range was because the datasets were measured with similar but different instruments and had spectral ranges that had been reduced by the suppliers of the original data. A key point to note is that all the data sets were pre-partitioned into calibration (60%) and test (40%) sets using the Kennard-Stone (KS) algorithm [51].

2.2. Data analysis

2.2.1. Preparing a multi-block data set from NIR spectra

A NIR data set for predictive modelling consist of spectra (X) of size $n \times p$, where n are the total number of samples and p are spectral variables. The response vectors (Y) are of size $n \times k$, where n are the total number of samples and k are total number of responses. To use the multi-block ensemble approaches, the NIR data can be processed with different pre-processings resulting in a multi-block data set. For example, if a NIR set of spectra (X), is pre-treated with six different pre-processings then the spectral data set will contain six data blocks, i.e., $[X_1, X_2, X_3, X_4, X_5, X_6]$, where $X_1 \ldots X_6$, are just differently pre-processed forms of the same spectra (X). In this study, to show the potential of the ROSA ensemble approach, the six different spectral sets were: raw data, 1st derivative, 2nd derivative, SNV, SNV +1st derivative, SNV +2nd derivative. The derivatives were calculated with the Savitzky-Golay algorithm [28] using a 2nd order polynomial and a window size of 13 points. Initially, all data sets were transformed to multi-block data using the same pre-processing combination and order. In a later part of the study, the effect of different pre-processing orders on the ensemble models was also explored.

2.2.2. Pre-processing ensembles with ROSA

ROSA [46] is a multi-block extension of the PLS technique and this study shows the potential of ROSA for the ensemble pre-processing for NIR spectra modelling. For a detailed description of the ROSA algorithm, readers are referred to the original paper [46]. In ROSA for ensemble pre-processing, the extraction of model components is organized as the competition between the covariance-maximizing candidate components computed from each differently pre-processed spectral block. At each step, the block component resulting in the smallest Y residual is assigned as the "winner" and taken to define the model component for that step. The competition between the next candidate components is constrained to the orthogonal complement of the subspaces spanned by the previous winning components. These constraints assure both orthogonal scores and loading weights. ROSA uses a forward selection approach of orthogonal components where blocks can be used several times to extract complementary information. A key trick of ROSA is that computationally intensive X deflations (common for classical PLS techniques) are replaced by faster Gram-Schmidt steps for computing the orthonormal scores and

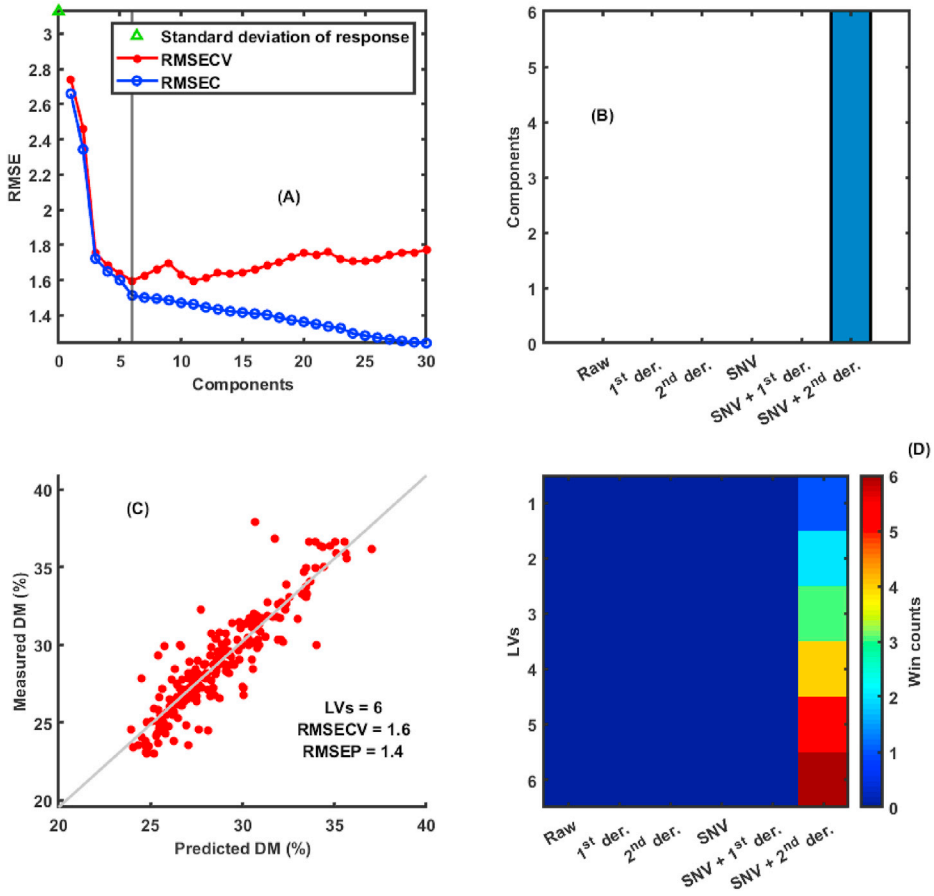

Fig. 5. Performance of PROSAC modelling on Olive data set. (A) Cross-validation plot for PROSAC, (B) LVs for each data block for PROSAC, (C) Prediction plot for PROSAC, and (D) Winning block order for LVs. In subplot (D), a change of colour along the vertical direction indicates that the latent variable is selected from that block. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

loading weights, thus making ROSA an extremely fast technique to explore many differently pre-processed data blocks. Furthermore, ROSA is also an order- and scale-independent technique, making it a perfect choice for the exploration of differently pre-processed data as pre-processing can sometimes change the data scales.

In this study, the ROSA decomposition was performed with the free code available in Ref. [46]. However, to optimize the ROSA components for each data sets, in this study a 10-fold Venetian blind cross-validation (CV) was integrated within ROSA. In the CV procedure, a range of components from 1 to 50 in step 1 was explored with cross-validation, and later, the elbow plots (RMSECV and RMSEC) were used to decide on the optimal number of components for the final ROSA model. The optimal number of components was decided by looking at the divergence of the RMSECV and RMSEC in the CV plots.

2.2.3. Baseline comparison

PROSAC was compared with the standard single-block PLS analysis on differently pre-processed data and with the recent multi-block ensemble pre-processing techniques SPORT [41] and PORTO [42]. In the results section the performance of PROSAC was first shown for all five data sets (Table 1) and compared with the standard PLS analysis. This was done to show the efficiency of PROSAC for ensemble modelling and to show how PROSAC is just a direct extension of single block PLS modelling. In the second part of the results, the performance of ROSA was compared with SPORT and PORTO using just the Mango data set. Comparison of the outcomes of the different multi-block techniques was based on the inspection of several aspects such as RMSEP, optimal number of latent variables, or the order of the blocks. Finally, the ability of PROSAC to model many pre-processed data blocks was also explored and the time requirements were reported.

The PLS analysis on single block data were performed using the *'plsregress'* function in MATLAB's 'Statistics and machine learning' toolbox and using a 10-fold CV as used for the PROSAC modelling. SPORT was implemented using the freely available MBA-GUI [52], and the number of latent variables for each block was optimized by exploring all combinations in the range of 0–10 to find the lowest RMSECV. PORTO modelling was done using the PO-PLS multi-block data analysis codes from NOFIMA (https://nofimamodeling.org/software-downloads-list/multiblock-regression-by-poso-pls/). After a preliminary data compression (which also provides noise-filtering) operated by retaining, for each block, the scores resulting by individual PLS modelling, several local cross-validation (CV) steps were performed in sequence, as discussed in Ref. [53], to find the optimal number of common and distinct components for building the PORTO model. All data analyses were performed using MATLAB (Release 2018b; The Mathworks, Natick, MA) on a workstation equipped with a NVidia GPU (GeForce RTX 2080 Ti), an Intel® Core™ i7-4770k @3.5 GHz and 64 Gb RAM, running Microsoft Windows 10 OS.

3. Results

3.1. Performance of PROSAC ensemble pre-processing for different data sets and comparison with single block PLS

This section presents the results of PROSAC analysis of five different data sets detailed in Table 1 and in block order of raw data, 1st derivative, 2nd derivative, SNV, SNV +1st derivative, SNV +2nd derivative. The comparison in this part is performed with the single block PLS analysis performed on each differently pre-processed data block. In

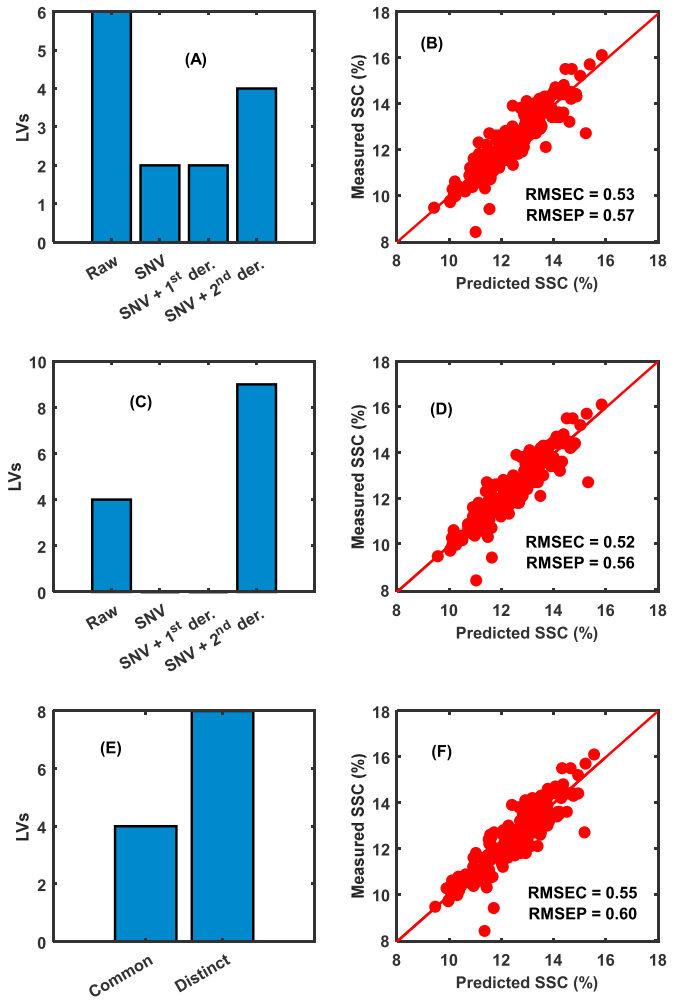

Fig. 6. Performance of PROSAC, SPORT and PORTO on Pear data set. (A) Winning block components for PROSAC, (B) Prediction plot for PROSAC, (C) Components from each block SPORT, (D) Prediction plot for SPORT, (E) Common and distinct components for PORTO, and (F) Prediction plot for PORTO.

Table 3
A summary of block orders used to compare its effect on PORTO, SPORT and PROSAC.

Block combination	Data block order
1	Raw, SNV, SNV +1st Derivative, SNV +2nd derivative
2	SNV, SNV +1st Derivative, SNV +2nd derivative, Raw
3	SNV +1st Derivative, SNV +2nd derivative, Raw, SNV
4	SNV +2nd derivative, Raw, SNV, SNV +1st derivative

Fig. 1, the analysis of the Pear data set is presented, where the cross-validation with PROSAC selected 14 components (Fig. 1A). Among the 14 components, many were extracted from 5 out of 6 of the pre-processed data blocks, showing that PROSAC learned an ensemble model (Fig. 1B). The trend of the block selection as a function of LVs is shown in Fig. 1D, where it can be noted that the initial winning block was the raw data block, and the other blocks start to contribute later for higher number of LVs. For the final optimal LVs = 14, there were 5 LVs from raw data block, 2 from 1st derivative, 3 from 2nd derivative, 2 from SNV and 2 from SNV followed by 1st derivative. The PROSAC model was tested on the independent test set giving an RMSEP = 0.56% (Fig. 1C). It can be noted that PROSAC based ensemble model gave the lowest RMSEP compared to all single block PLS regressions performed on individual data blocks with a similar number of components (Table 2). Similarly, for the Avocado data set (Fig. 2), 9 out of 10 LVs were extracted from the SNV+2nd derivative data leading to a RMSEP = 1.22%. For the Avocado data set, the RMSEP was the same as the PLS analysis on the SNV+2nd derivative data with an equal number of components, i.e., 10. With PROSAC modelling there was a slightly lower difference between the RMSEC and RMSEP compared to PLS analysis on the SNV+2nd derivative data which may indicate the PROSAC models may generalise well, however, the difference was too low to justify the benefits of PROSAC further.

For the Apple data set (Fig. 3), as with the Pear data set, PROSAC gave an ensemble model selecting components from 4 out of 6 data blocks. Since the model reached performance comparable to that of PLS on the individual blocks, the main contribution of PROSAC here can be considered as simply saving time by developing a single PROSAC model compared to developing several single block PLS models to find the best pre-processing. For the Mango data set, PROSAC built an ensemble model by selecting 5 out of 6 data blocks. The PROSAC model (Fig. 4C) for the Mango data set performs better than the models made on normalised data, and a combination of normalisation followed by derivative (Table 2). However, the PROSAC model performed more poorly than the PLS model built solely on derivative pre-processed data (Table 2). In Table 2, it can be noted that the normalisation deteriorates the model. Looking at the order of the selected blocks per component (Fig. 4D), it can be noted that the PROSAC model for Mango selected the first component from the SNV normalised data block. Such a selection of the first component from the normalised data block makes it impossible, due to the block deflation step in ROSA, to retrieve a good model as the non-normalised space gets eliminated during the deflation. Hence, a reason for the failure of the PROSAC model here may be found in the limitations associated to the heuristic of the ROSA algorithm which does not allow already selected model components to be updated during the selection of future model components. Basically, the ROSA model heuristic involves a forward stepwise selection but without any backward elimination step to rejudge the selection of the model components. Currently, there is room for improvement of the ROSA algorithm as its current heuristic for model component selection may in some cases lead to poor models, although this was observed for only 1 out of the 5 data sets in this study.

In the case of the Olive (Fig. 5) data set, PROSAC modelling selected just a single data block, i.e., SNV +2nd derivative. Such a selection of a single data block suggests that PROSAC modelling does not only perform ensemble modelling but can also identify an individual pre-processing as

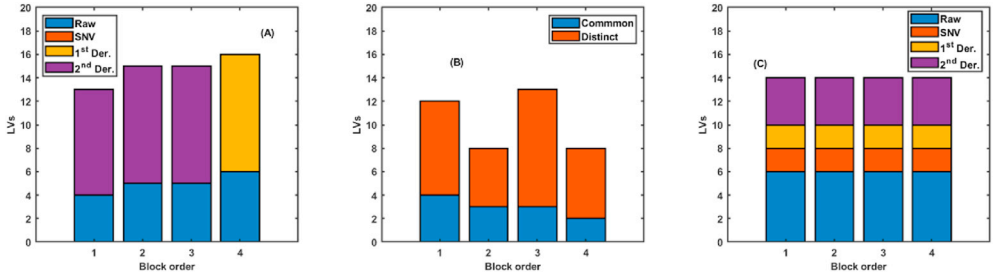

Fig. 7. Effect of changing block orders on (A) SPORT, (B) PORTO, and (C) PROSAC. The analyses were carried out on the Pear data set.

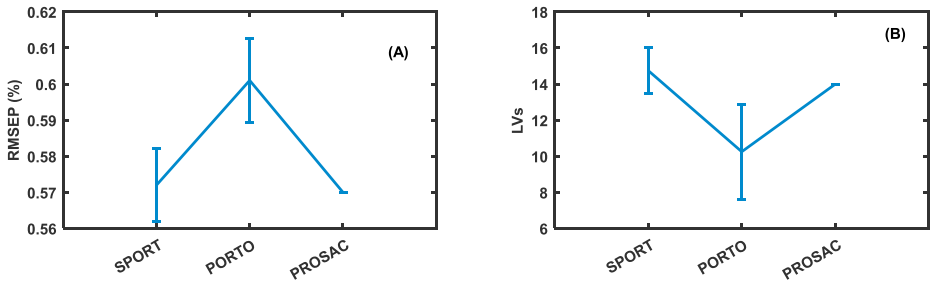

Fig. 8. Effect of changing block combinations for SPORT, PORTO and PROSAC on (A) RMSEP, and (B) LVs. The analysis was carried out on Pear data set.

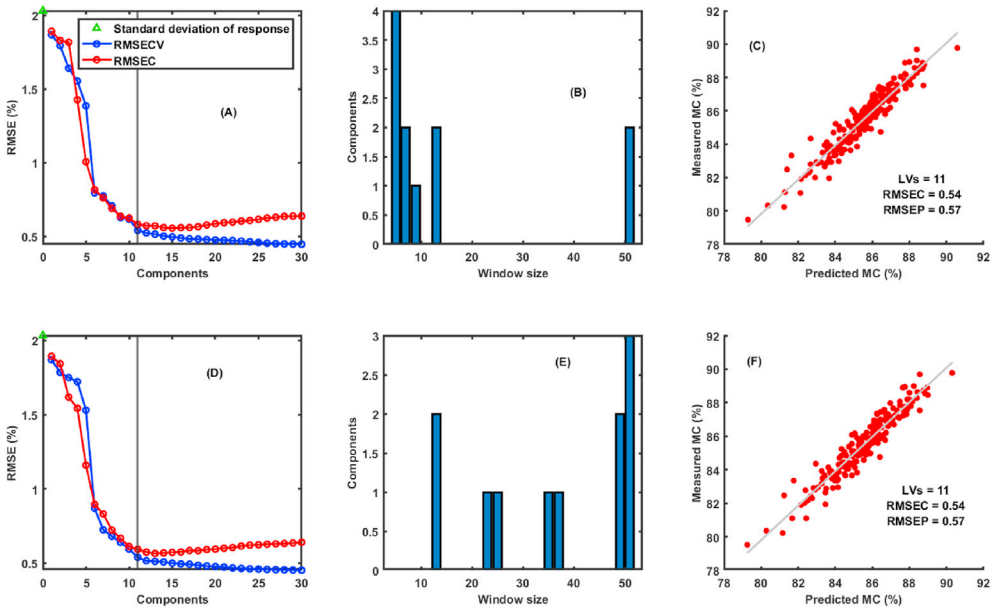

Fig. 9. Exploration of the optimal window size for Savitzky-Golay 1st (top row) and 2nd derivative (bottom row) with PROSAC. (A) Cross-validation plot for selecting optimal PROSAC components for 1st derivative data, (B) Winning blocks, (C) Prediction plot for PROSAC on 1st derivative data, (D) Cross-validation plot for selecting optimal PROSAC components for 2nd derivative data, (E) Winning blocks, and (F) Prediction plot for PROSAC on 2nd derivative data. The analysis was carried out on Mango data set.

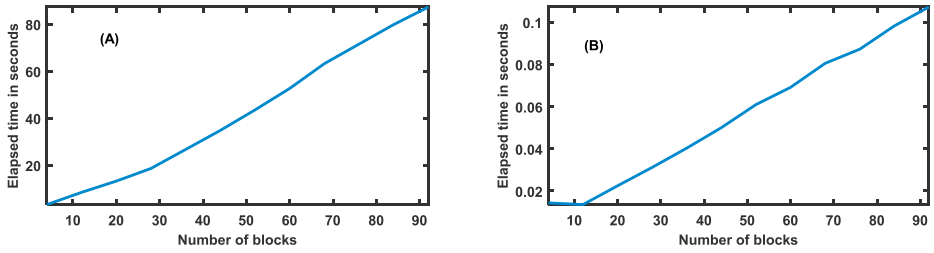

Fig. 10. A comparison of elapsed time for PROSAC analysis for ensemble pre-processing as a function of increasing number of pre-processing blocks. (A) Time for cross validating the model for tuning optimal components, and (B) Time for PROSAC model development and testing on a new data set. The analysis was carried out on the Mango data set.

the best one, when only a single pre-processing is sufficient to achieve the optimal model. For the Olive (Fig. 5) data set, it can be noted that, as to be expected, the performance of the PROSAC modelling was exactly like that of the corresponding single block PLS analysis (Table 2).

3.2. A comparison of ensemble pre-processing using ROSA with SPORT and PORTO

The performance of PROSAC was also benchmarked with the two popular multi-block ensemble techniques SPORT and PORTO. The analysis was carried out on the Pear data set and on four data blocks in the order: Raw, SNV, SNV +1st Derivative, SNV +2nd derivative. The analysis was reduced to one data set and from six to four blocks due to the high optimization time cost for SPORT and PORTO. The effect of block order on SPORT, PORTO and PROSAC is examined later in this study. In Fig. 6, it can be noted that SPORT achieved the lowest RMSEP, and outperformed PROSAC and PORTO, even if a close look at the model statistics shows that the differences in RMSEP's of SPORT, PORTO and PROSAC were minimal. This ranking is quite logical, since SPORT comprehensively searches for the best LV combination, whilst PROSAC uses a stepwise heuristic, and can be affected by local minima. In this case, the key benefit of the PROSAC approach compared to SPORT and PORTO is that it is independent of the block order. This is important for ensemble pre-processing modelling as the user does not need to worry about arranging the blocks in an optimal order. To show the effect of changing block order on PORTO, SPORT and PROSAC analyses, the analysis presented in Fig. 6 was repeated for different block order combinations as explained in Table 2.

The results of the SPORT, PORTO and PROSAC analyses performed on the different block orders (Table 3) are presented in Fig. 7 and Fig. 8. It can be noted that with change in block order, the number of components corresponding to the different data blocks changes for both SPORT (Fig. 7A) and PORTO (Fig. 7B). In the case of PORTO, it is the total number of common and distinct components that changes. On the other hand, for PROSAC (Fig. 7C), the number of components from the blocks are always the same and for the same data blocks. Such a uniform selection of components and blocks by PROSAC shows that it is indeed an order-independent technique, which can be of great use for ensemble pre-processing modelling. The effect of such changes in block order can also be seen in the RMSEP of the SPORT and PORTO models (Fig. 8A), while the RMSEP for the PROSAC stayed the same. Please note that the good performance of SPORT and PROSAC than the PORTO could be related to the number of LVs extracted by SPORT and PROSAC being higher (Fig. 8B). Again, PROSAC maintained its performance despite the changing block orders.

3.3. Further comments on PROSAC for pre-processing exploration

In the earlier section, the potential of PROSAC for ensemble modelling and its comparison with other multi-block ensemble approaches was presented. In this section the capabilities of PROSAC to handle several blocks are highlighted. To show this, a ROSA model was used for 24 data blocks (Mango data set) containing the Savitzky-Golay derivative over the window ranges of from 5 to 51 in steps of 2. This means using PROSAC to explore 24 blocks. The results for the 1st and the 2nd derivatives are shown as the top and bottom rows in Fig. 9. It can be noted that out of the wide range of intervals explored, only a subset of window intervals was found to be the most explanatory. Learning from such an analysis can allow the selection of a subset of the window sizes for further exploration in combination with other pre-processings such as normalisation or scatter corrections.

As noted in the analysis presented in Fig. 9, a total of 24 data blocks were modelled with PROSAC to evaluate the possible interplay of different window sizes for the SavGol derivatives. Handling such a substantial number of blocks is challenging with the traditional methods such as SPORT and PORTO which rely on a more complex heuristic, due to the sequential or parallel nature of the analyses. To have an insight into the time requirements, plots of the computational cost as a function of the number of data blocks are shown in Fig. 10. It can be noted that with increasing number of blocks, the model optimization (10-fold CV) time increased; however, optimising PROSAC on the largest block size of 96, took just ~85 s. The model selection process in this case involved exploring a total of $10 \times 50 \times 96 = 48000$ total candidate components for data of dimensions 300×103. Similarly, time requirement for developing and testing a final PROSAC model based on 96 data blocks was ~0.1 s. Such fast handling of a substantial number of blocks makes PROSAC a unique tool where many pre-processings can be handled in a single run, which seems currently challenging for other techniques like SPORT and PORTO.

4. Conclusions

This study proposed the PROSAC multi-block analysis approach for ensemble pre-processing modelling of NIR spectral data. The analysis of five different real NIR data sets showed that in four out of five data sets, PROSAC ensemble modelling either achieved better performance by using an ensemble of information from differently pre-processed NIR data or achieved performance like PLS analyses performed on individually pre-processed NIR data. A key point to note is that the single block PLS requires training several models for each pre-processing, while for PROSAC it is done in a single run, thus saving a substantial amount of time for model exploration. Furthermore, PROSAC converges to a single block PLS analysis when only a single block of data is selected as the winning block. In such a way, PROSAC chooses the optimal pre-processing rather than an ensemble of pre-processings.

PROSAC, PORTO, and SPORT are all ensemble learning methods based on multiple blocks. They differ primarily in the heuristics they use to explore how to combine the latent variables extracted from the blocks. Because they use different heuristics, they produce different results and have different advantages and disadvantages. On one data set, PROSAC modelling did not perform as well as some single block models. The reason for this result is related to the heuristic used by PROSAC, based on forward stepwise selection. In some cases, if a particular component is selected first for the model, the subsequent deflation phase could alter the data set and compromise the rest of the selections. This shows that the PROSAC algorithm could be further improved by adding a backward step to re-evaluate the usefulness of previously selected model components. Compared to SPORT and PORTO, PROSAC was unaffected by the order in which the pre-processing blocks were arranged. Furthermore, the test of PROSAC on a substantial number of data blocks showed that it is a fast approach to perform a multi-block ensemble pre-processing. It was found that optimising a 96 data block PROSAC model took ~80 s, which is difficult to achieve with any other multi-block ensemble pre-processing approach. However, it should be stressed that the speed of PROSAC will always depend on the implementation and data dimensions, as with any PLS based algorithm. The ability of PROSAC to perform fast order- and scale-independent ensemble pre-processing and converge to a traditional PLS when only a single pre-processing is sufficient, makes it a useful tool for NIR calibration. Note that PROSAC is an application of the ROSA multi-block modelling approach which, in turn, is an extension of PLS regression, and therefore, PROSAC supplies all the relevant parameters for model interpretability, such as regression coefficients, scores and loadings.

On the other hand, it should be stressed that, in principle, when selecting components from multiple versions of the same dataset, there could always be the risk of overfitting. However, due to its algorithmic nature, this is not the case for PROSAC (as well as for SPORT or PORTO), as is shown by the closeness of the values of the RMSEC, RMSECV and RMSEP in all the reported examples.

Based on the findings from this study, it can be concluded that ROSA ensemble pre-processing gives promising directions to end the era of exhaustive pre-processing search and optimization for modelling NIR data.

Author statement

Puneet Mishra: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft, **Jean Michel Roger:** Conceptualization, Methodology, Writing - Original Draft, **Federico Marini:** Conceptualization, Methodology, Writing - Original Draft, **Alessandra Biancolillo:** Conceptualization, Methodology, Writing - Original Draft, **Douglas N. Rutledge:** Conceptualization, Methodology, Writing - Original Draft.

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.

References

[1] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[2] B.G. Osborne, Near-infrared spectroscopy in food analysis, in: R.A. Meyers, R.J. McGorrin (Eds.), Encyclopedia of Analytical Chemistry, John Wiley and Sons, NY, 2006, https://doi.org/10.1002/9780470027318.a1018.
[3] R. Bro, Multivariate calibration: what is in chemometrics for the analytical chemist? Anal. Chim. Acta 500 (2003) 185–194.
[4] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochim. Acta B Atom Spectrosc. 58 (2003) 767–782.

[5] P. Mishra, S. Lohumi, H. Ahmad Khan, A. Nordon, Close-range hyperspectral imaging of whole plants for digital phenotyping: recent applications and illumination correction approaches, Comput. Electron. Agric. 178 (2020) 105780.
[6] P. Mishra, A. Nordon, J.-M. Roger, Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques, J. Pharmaceut. Biomed. Anal. (2020) 113684.
[7] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, Trends Food Sci. Technol. 18 (2007) 590–598.
[8] J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, Anal. Chim. Acta 896 (2015) 34–51.
[9] K.B. Walsh, V.A. McGlone, D.H. Han, The uses of near infra-red spectroscopy in postharvest decision support: a review, Postharvest Biol. Technol. 163 (2020) 111139.
[10] K.B. Walsh, J. Blasco, M. Zude-Sasse, X. Sun, Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: the science behind three decades of commercial use, Postharvest Biol. Technol. 168 (2020) 111246.
[11] R.A. Crocombe, Portable spectroscopy, Appl. Spectrosc. 72 (2018) 1701–1751.
[12] N. Prieto, O. Pawluczyk, M.E.R. Dugan, J.L. Aalhus, A review of the principles and applications of near-infrared spectroscopy to characterize meat, fat, and meat products, Appl. Spectrosc. 71 (2017) 1403–1426.
[13] B. Stenberg, R.A. Viscarra Rossel, A.M. Mouazen, J. Wetterlind, D.L. Sparks, Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science, Advances in Agronomy, Academic Press, NY, 2010, pp. 163–215.
[14] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. (2019) 158.
[15] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS method, in: B. Kågström, A. Ruhe (Eds.), Matrix Pencils. Lecture Notes in Mathematics, first ed., Springer, Berlin/Heidelberg, Germany, 1983, pp. 286–293.
[16] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.
[17] P. Mishra, D. Passos, Realizing transfer learning for updating deep learning models of spectral data to be used in a new scenario, Chemometr. Intell. Lab. Syst. 212 (2021) 104283.
[18] P. Mishra, D. Passos, A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit, Chemometr. Intell. Lab. Syst. 212 (2021) 104287.
[19] P. Mishra, D. Passos, Deep multiblock predictive modelling using parallel input convolutional neural networks, Anal. Chim. Acta 1163 (2021) 338520.
[20] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac. Trends Anal. Chem. 132 (2020) 116045.
[21] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trac. Trends Anal. Chem. 28 (2009) 1201–1222.
[22] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing methods, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, 2 nd ed., Elsevier, Oxford, UK, 2020, pp. 1–75.
[23] R.F. Lu, R. Van Beers, W. Saeys, C.Y. Li, H.Y. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020) 111003.
[24] P. Mishra, D.N. Rutledge, J.-M. Roger, K. Wali, H.A. Khan, Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction, Talanta 229 (2021) 122303.
[25] P. Mishra, J.-M. Roger, D.N. Rutledge, A short note on achieving similar performance to deep learning with practical chemometrics, Chemometr. Intell. Lab. Syst. 214 (2021) 104336.
[26] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[27] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.
[28] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[29] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (2003) 1363–1367.
[30] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[31] W. Windig, J. Shaver, R. Bro, Loopy MSC: a simple way to improve multiplicative scatter correction, Appl. Spectrosc. 62 (2008) 1153–1159.
[32] A. Kohler, J.H. Solheim, V. Tafintseva, B. Zimmermann, V. Shapaval, 3.03 - model-based pre-processing in vibrational spectroscopy, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, second ed., Elsevier, Oxford, 2020, pp. 83–100.
[33] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020) 111271.
[34] J. Torniainen, I.O. Afara, M. Prakash, J.K. Sarin, L. Stenroth, J. Toyras, Open-source python module for automated preprocessing of near infrared spectroscopic data, Anal. Chim. Acta 1108 (2020) 1–9.
[35] K.H. Liland, T. Almøy, B.-H. Mevik, Optimal choice of baseline correction for multivariate calibration of spectra, Appl. Spectrosc. 64 (2010) 1007–1016.
[36] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[37] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50 (2013) 96–106.
[38] P. Stefansson, K.H. Liland, T. Thiis, I. Burud, Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations, J. Chemometr. 34 (2020), e3195.
[39] P. Mishra, T. Verkleij, R. Klont, Improved prediction of minced pork meat chemical properties with near-infrared spectroscopy by a fusion of scatter-correction techniques, Infrared Phys. Technol. 113 (2021) 103643.
[40] P. Mishra, S. Lohumi, Improved prediction of protein content in wheat kernels with a fusion of scatter correction methods in NIR data modelling, Biosyst. Eng. 203 (2021) 93–97.
[41] P. Mishra, F. Marini, A. Biancolillo, J.-M. Roger, Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta 223 (2020) 121693.
[42] L. Xu, Y.-P. Zhou, L.-J. Tang, H.-L. Wu, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta 616 (2008) 138–143.
[43] X. Bian, K. Wang, E. Tan, P. Diwu, F. Zhang, Y. Guo, A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples, Chemometr. Intell. Lab. Syst. 197 (2020) 103916.
[44] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 199 (2020) 103975.
[45] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy, Chemometr. Intell. Lab. Syst. 212 (2020) 104190.
[46] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, J. Chemometr. 30 (2016) 651–662.
[47] P. Mishra, D. Passos, Deep chemometrics: validation and transfer of a global deep near-infrared fruit model to use it on a new portable instrument, J. Chemometr. 35 (2021) e3367.
[48] P. Mishra, E. Woltering, Handling batch-to-batch variability in portable spectroscopy of fresh fruit with minimal parameter adjustment, Anal. Chim. Acta 1177 (2021) 338771.
[49] S.L. Teh, J.L. Coggins, S.A. Kostick, K.M. Evans, Location, year, and tree age impact NIR-based postharvest prediction of dry matter concentration for 58 apple accessions, Postharvest Biol. Technol. 166 (2020) 111125.
[50] X. Sun, P. Subedi, R. Walker, K.B. Walsh, NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment, Postharvest Biol. Technol. 163 (2020) 111140.
[51] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[52] P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI, A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing, Chemometr. Intell. Lab. Syst. 205 (2020) 104139.
[53] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, Food Qual. Prefer. 24 (2012) 8–16.