Postharvest Biology and Technology 168 (2020) 111271

Contents lists available at ScienceDirect

Postharvest Biology and Technology

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

SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials

Puneet Mishra[a,*], Jean Michel Roger[b,c], Douglas N. Rutledge[d,e], Ernst Woltering[a,f]

[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] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005, Paris, France
[e] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia
[f] Horticulture and Product Physiology Group, Wageningen University, Droevendaalsesteeg 1, P.O. Box 630, 6700AP, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
SOPLS
Multi-block
Chemometric
Data-fusion
Scatter correction

ABSTRACT

Near-infrared spectroscopy (NIRS) is a key non-destructive technique for rapid assessment of the chemical properties of food materials. However, a major challenge with NIRS is the mixed physicochemical phenomena captured by the interaction of the light with the matter. The interaction often results in both absorption and scattering of the light. The overall NIRS signal therefore contains information related to the two phenomena mixed. To predict chemical properties such as dry matter, Brix and lipids, light refelction/absorption is used. Therefore, when the aim of the data analysis is to predict chemical components, it is necessary to remove as much as possible the scattering effects from the spectra. Several pre-processing techniques are available to do this, but it is often difficult to decide which one to choose. In this article we present the use of a recently developed pre-processing approach, sequential pre-processing through orthogonalization (SPORT), to improve the predictive power of multivariate models based on NIR spectra of food materials. The SPORT approach utilizes sequential orthogonalized partial least square regression (SOPLS) for the fusion of data blocks corresponding to several spectral preprocessing techniques. The results were compared with commonly used pre-processing techniques in the analysis of food materials by NIRS. The comparison was made by analyzing 5 different datasets comprised of apples, apricots, olive oils and grapes associated with chemical properties such as dry matter (DM), Brix, lipids and citric acid. The datasets were from both reflection and transmission measurements. The results showed that the fusion-based pre-processing methodology is an ideal choice for pre-processing of NIRS data. For four out of five datasets, the prediction accuracies (high R^2_{pred} and low RMSEP) were improved. The improvement led to as much as a 20 % increase in R^2_{pred} and a 25 % decrease in RMSEP compared to the standard 2nd derivative pre-processing. The pre-processing fusion was more effective for the reflection mode compared to the transmission mode. Multiple pre-processing techniques provided complementary information, and therefore, their fusion using the SPORT approach improved the model performance. The methodology is not only applicable to food materials but can in fact be used as a general pre-processing approach for all types of modeling of spectral data.

1. Introduction

Near-infrared spectroscopy (NIRS) is the key technique for non-destructive exploration of food materials (Nicolai et al., 2007; Kamruzzaman et al., 2015; Wang et al., 2015; Arendse et al., 2018; Walsh et al., 2020). The technique utilizes the interaction of infrared radiation with matter and captures both the absorption and the scattering characteristics. Depending on the physical state of the samples (such as solid or liquid), NIRS can be deployed in either transmission or reflection modes (Pasquini, 2018). The transmission is the preferred mode for liquid samples (Armenta et al., 2010; Gómez-Caravaca et al., 2016) and rarely implemented for solid samples unless they are very thin (Sierra et al., 2008), unless, the aim is to detect internal disorders in fresh fruits. Sometimes for solid samples such as fruits, the partial-transmission can be utilized which involves excitation with IR radiation and recording the response at different points on the sample (Nicolai

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

https://doi.org/10.1016/j.postharvbio.2020.111271
Received 14 May 2020; Received in revised form 7 June 2020; Accepted 9 June 2020

et al., 2007; Lu et al., 2020). These different sample presentation modes result in different attenuation in the signals and can capture complementary information.

Food materials such as fruits are measured in the reflection mode to predict properties such as dry matter percentage (DM), titratable acidity (TA) and Brix (total soluble solids). These parameters can indicate the maturity and ripening stage of fruits. However, the implementation of NIRS for fruit comes with a challenge since the recorded spectrum contains a mixture of light absorption and light scattering effects (Lu et al., 2020). The light absorption is due to the presence of chemical components while the light scattering is mainly due to the physical structure of the fruit peel and flesh. The presence of scattering in the signal affects the predictive modelling, leading to lower accuracy and poor performance (Saeys et al., 2019), since in order to predict the chemical components, the main phenomena in the NIRS to be taken into account are the absorption characteristics (Lammertyn et al., 2000). Therefore, it is common practice to correct for the scattering effects in the NIRS spectra before data modelling (Martens et al., 2003; Rinnan et al., 2009). The effects of scattering are not only present in solid foods as similar effects can be found in liquid foods such as oil measured in the transmission mode. In the case of transmission, the scattering effects depend on both the concentration of analytes in the liquid and the path length. In summary, the NIRS spectra (acquired in any mode) can be affected by a range of scattering effects which are not easy to correct for using physical, mechanistic models. On the other hand, scatter correction using empirical mathematical methods of pre-processing is easy and fast.

Pre-processing is a key step in NIRS modelling, and several methods are available (Rinnan et al., 2009). The first step of pre-processing in NIRS includes visualization of the data and removal of extreme bands which are dominated by noise. The second step involves performing window-based smoothing operations to remove any high-frequency noise. A common method to do this is based on the Savitzky-Golay smoothing algorithm (SAVGOL) which involves fitting a polynomial of chosen order within a band of specified size, which is moved across the complete spectrum (Rinnan et al., 2009). Ideally, in the presence of only absorption features, the smoothed spectra should be ready for regression or classification modeling. However, due to the dominance of scattering effects, the smoothing step is usually followed by scattering correction methods, several of which are available in the chemometric literature. The most common methods are the estimation of 2nd derivative of the spectra as it can easily remove first-order additive (baseline shift) effects and also reveal underlying peaks that otherwise may not be visible (Rinnan et al., 2009). Another commonly used method is the standard normal variate (SNV) and involves treating each spectrum by subtraction of its mean spectral intensity from each intensity response and then division by its standard deviation estimated over the spectral domain (Barnes et al., 1989). Doing SNV can remove additive and multiplicative effects. Both the 2nd derivative and SNV prove to be of high value in NIRS modeling and usually improve the model predictive performance. Another popular approach is multiplicative scatter correction (MSC), where it is assumed that the spectrum consists of a multiplicative, an additive and a residual part (Isaksson and Næs, 1988). Extended MSC (EMSC) further takes into account higher order complex relations to model these effects (Martens et al., 2003). Improvements and alternatives to SNV have also been proposed such as robust normal variate (RNV) (Guo et al., 1999), probabilistic quotient normalization (PQN) (Dieterle, et al. 2006) and variable sorting for normalization (VSN) (Rabatel et al., 2020). Other variants locally apply pre-processing by dividing the spectra into multiple chunks (Bi et al., 2016). In summary, there are many pre-processing methods available in the chemometric domain to remove/reduce the scattering effects from NIRS spectra. A summary of available pre-processing methods can be found in (Roger et al., 2020b).

Choosing the right pre-processing is always a challenge. Current approaches to perform pre-processing are not optimal and the need to have a proper pre-processing strategy has been highlighted (Engel et al., 2013). With that in mind, a design of experiments (DOE) -based approach was proposed in (Gerretzen et al., 2015) where a combination of pre-processing methods was selected by evaluating model performances in relation to ordered pre-processing steps. The pre-processing strategy was broadly divided into four steps and in sequential order, such as baseline correction, scatter correction, smoothing and scaling. In this way, the methodology should be able to find a good combination of pre-processing methods in an order, but never were two or more scattering correction techniques used in a complementary way. Different scatter correction methods may enhance the NIRS spectra in different ways and thus be complementary. This hypothesis was demonstrated in (Xu et al., 2008) where different pre-processing methods were used in an ensemble approach which led to the conclusion that an ensemble of spectral data pre-processed with techniques resulted in more stable and accurate models. Another ensemble approach showed that selective pre-processing can give better results compared to traditional approaches (Bian et al., 2020). In summary, using complementary information from multiple scatter correction techniques, rather than just using only one, can be beneficial.

In recent years, multi-block analysis has been gaining attention for performing data fusion in chemometrics (Smilde et al., 2017; Måge et al., 2019; Song et al., 2020). Especially useful are the partial least square (PLS) -based multi-block regression methods that can search for the complementary information in different blocks (Biancolillo et al., 2016, 2017; Biancolillo et al., 2019). One such method is sequential orthogonalized PLS (SOPLS) which performs PLS decomposition sequentially on the blocks to extract from each the variability that is related to the response variables. The main benefit of the sequential approach is that complementary information is extracted from each block. Furthermore, it allows selection of blocks which contribute to the improvement of the model. SOPLS has already been widely used for data fusion from multiple sensors (Niimi et al., 2018; Awhangbo et al., 2020; Firmani et al., 2020), but it could also be used to combine the results of applying multiple preprocessing techniques to a single set of data. Such an approach, called sequential pre-processing through orthogonalization (SPORT), was recently developed to sequentially extract information from different blocks, corresponding to different pre-processing, so that the maximum of the variation in the response variables is explained (Roger et al., 2020a). In the present work, we hypothesized that several scatter correction techniques when used in a complementary way can lead to better predictive performance of NIRS models.

In particular, the present work aims to demonstrate the usefulness of the SPORT pre-processing (Roger et al., 2020a) methodology to fuse data from different scatter correction techniques. The methodology is tested on five different datasets related to the prediction of chemical components in food materials. Properties such as DM, Brix and lipids were predicted in fruit and oils. The 2nd derivative and a combination of SNV + 2nd derivative pre-processing was used as the reference pretreatment method to which the performance of SPORT-based pre-processing fusion was compared.

2. Material and methods

2.1. Datasets

The 5 datasets included both products with complex matrices such as fresh fruit, and liquid products such as oils. More details on each dataset are provided in separate sections. A summary of dataset size and the reference measurement ranges is provided in Table 1. The calibration (70 %) and the test set (30 %) were partitioned using the Kennard Stone algorithm (Kennard and Stone, 1969).

2.1.1. Apple data

The apple data contain 625 individual fruits and was same as used in the literature (Roger et al., 2003). As a reference, Brix measurement

Table 1
A summary of all the datasets used in the study.

Material	Samples	Spectral range	Fruit property/reference	Reference range	Measurement mode
Apple	625	499 – 1018 nm	SSC	7.5–18 %	Reflection
Grape	245	499 – 1018 nm	SSC	7 – 27 %	Transmission
Olive oil	187	1000 – 2222 nm	Fats (Linoleic acid (LOL)) component rate from gas chromatography	0.1–2.2 %	Transmission
Olive fruit	535	669 – 1122 nm	DM	21.97 – 37.19 %	Reflection
Apricot	662	800 – 2772 nm	Citric acids	3.4–44 mmol kg^{-1}	Reflection

was performed by sampling the juice at the exact same location where the NIRS measurement was performed. The dataset was obtained from the website of the ChemHouse project (https://chemproject.org).

2.1.2. Grape data

The Grape data contain spectra from 245 individual grape berries and the corresponding Brix measurements were done by extracting the juice from the grape. The dataset was obtained from the website of the ChemHouse project (https://chemproject.org).

2.1.3. Olive oil data

The spectra from 187 olive oil samples and reference were the triglyceride measurements (Galtier et al., 2007). The dataset was obtained from the website of the ChemHouse project (https://chemproject.org).

2.1.4. Olive fruit data

The olive fruit dataset contain spectra of 535 olive fruits and corresponding reference DM measurements. More details on the data acquisition and experimental protocol can be found in (Sun et al., 2020).

2.1.5. Apricot data

The apricot dataset consisted of FT-IR spectra of 662 apricots and corresponding reference was citric acid content. More details on the dataset can be found in (Bureau et al., 2009). The dataset was obtained from the website of the ChemHouse project (https://chemproject.org).

2.2. Data analysis

The data analysis involved the application of different pre-processing techniques and using PLS and SOPLS regression. Given that the samples were highly scattering materials, 4 different scatter correction pre-processing techniques were preselected to use with the SPORT pre-processing procedure. The four scatter correction approaches were multiplicative scatter correction (MSC) using the mean as the reference measurement, variable sorting for normalization (VSN), standard normal variate (SNV) and 2nd derivative.

2.2.1. Pre-processing

2.2.1.1. Multiplicative scatter correction

MSC: MSC is a common scatter removal technique used for correction of both additive and multiplicative effects (Isaksson and Næs, 1988; Martens et al., 2003). MSC models the spectra as a mixture of scattering and absorbance. It uses a reference spectrum (often mean) and tries to rotate all the other spectra so that they are as close as possible to the reference spectrum; by changing the scale and offset of the spectra. The MSC assumes that the diffusion scattering is the same for all samples and at all wavelengths. To estimate the slope and offset terms for correction of the spectra; MSC utilizes a least-squares regression. Once the slope and offset terms are estimated; MSC performs the correction for each individual spectrum as in Eq. 1.

$$x_{corr} = (x\text{-}a)/m \quad (1)$$

where, $\mathbf{x_{corr}}$ is the corrected spectrum, $\mathbf{x}$ is the raw spectrum, $\mathbf{a}$ is the offset parameter extracted by MSC and $\mathbf{m}$ is the extracted slope parameter. All the spectra are corrected with the same parameters as Eq. 1. In the present work, MSC was implemented using the codes presented in (Roger et al., 2020b).

2.2.1.2. Standard normal variate

Standard normal variate (SNV) is a popular technique used for normalizing NIR spectra with the aim of reducing the multiplicative effects due to light scattering and additive effects presented as differences in global signal intensities (Barnes et al., 1989). The method is a simple calculation of the z-score, but the difference is that the z-scoring is performed on each spectrum, instead of on each variable. The method does not require any external parameter estimation as does MSC, since the correction parameters are the mean and the standard deviation of each spectrum. During the correction, the mean is subtracted from the spectrum (all wavelengths) and then each point in the spectrum is divided by the spectrum standard deviation. Subtraction of the mean (of a given spectrum) is a baseline correction and the division by the standard deviation reduces differences in global intensities. However, SNV has the defect of mathematical closure due to the division by the standard deviation of the whole spectrum, and this means that changes in the shape of one part of the spectrum can have an effect over the entire spectrum. Such a distortion can influence the robustness of the models as the b-coefficients no longer correspondent to the physical reality present in the NIRS spectra. In the present work, the SNV was implemented using the MATLAB codes explained in literature (Roger et al., 2020b).

2.2.1.3. Variable sorting for normalization

Variable sorting for normalization (VSN) is a recently developed scatter correction technique which compensates for the drawbacks of both SNV and MSC (Rabatel et al., 2020). In comparison to the SNV and the MSC, the VSN extracts weighted global statistics by assuming that not all the bands are affected equally the scattering effects. The estimated weights from the VSN can be integrated to the SNV or MSC global statistics to perform a weighted SNV or a weighted MSC. The weights are estimated based on a RANSAC algorithm which determines to what extent a wavelength is affected by pure additive and multiplicative effects. In this way, variables that are strongly related to chemical components have a low weight and have negligible effect in the calculation of the parameter to be used to correct for the global intensity effect. The VSN approach does not require a reference spectrum to perform the estimation. In the present work, the VSN was implemented as presented in (Rabatel et al., 2020; Roger et al., 2020b).

2.2.1.4. 2nd derivative

The second derivative can remove both the linear baseline slope variations and additive effects. The derivative in the spectral domain is done with the help of a polynomial fitting within a window of chosen size (to have fair comparison in the present work a default of 15 was used for all datasets) and then the differential of the polynomial is calculated. The algorithm most often used for carrying out this operation is the Savitzky-Golay (SAVGOL) algorithm. In the present work, the SAVGOL algorithm was implemented using the MATLAB codes explained in literature (Roger et al., 2020b).

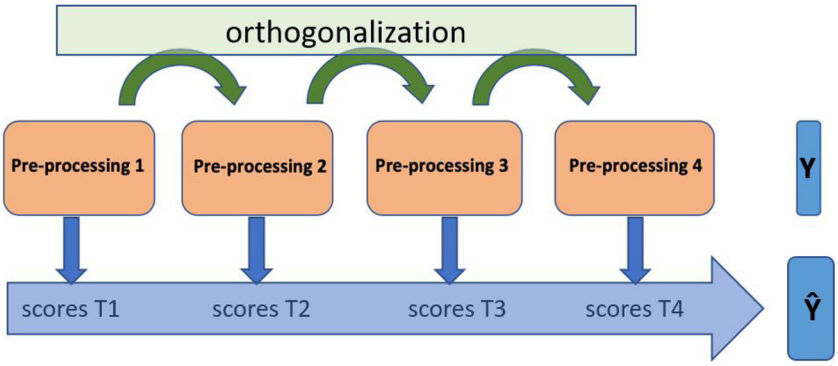

Fig. 1. A schematic of the SPORT pre-processing approach. T1, T2, T3 and T4 explains the scores extracted by SPORT from each block. Y is the response and Ŷ is the predicted response.

2.4. SPORT approach to pre-processing

Sequential pre-processing through orthogonalization (SPORT) is a recent approach to combine multiple pre-processing techniques to extract complementary information (Roger et al., 2020a). The complementary information from each pre-processed dataset is extracted and combined to improve model accuracies. The main core of the SPORT pre-processing is sequential orthogonalized partial-least square (SOPLS). SOPLS belongs to the family of multi-block PLS methods where, in addition to the PLS regression, an orthogonalization step is applied while calibrating the models with different data matrices (Biancolillo et al., 2019). SOPLS involves extraction of information sequentially from each measurement block used to build the calibration model, i.e. the aim is to incorporate blocks of data one at a time and to assess the incremental and additional contributions. A novel use of multiblock methods can also be interpreted as a boosting of different pre-processing techniques as shown in Fig. 1. Boosting with SPORT is a new approach to pre-processing and has proven to increase the prediction accuracies of the models. Firstly, a PLSR model is fitted between the **Y** and the first pre-processing block. Then based on the scores of the first blocks obtained with the PLSR the second block is orthogonalized. The orthogonalized second block is then fitted on the residuals of **Y** and so on it continues for as many blocks. The algorithm for a two pre-processing blocks ($\mathbf{X_1}$ and $\mathbf{X_2}$) as presented in (Roger et al., 2020a) is as follow:

1 The **Y** responses are fitted to the $\mathbf{X_1}$ with the PLSR and scores **T1** are obtained
2 $\mathbf{X_2}$ is orthogonalized with the scores of obtained from the first PLSR **(T1)**
3 The orthogonalized $\mathbf{X_2}$ is used to predict the **Y** residuals
4 Step 1, 2, 3 can be repeated for **p** number of blocks corresponding to **p** different pre-processing techniques
5 The final model is obtained by summing up all the PLSR models

The LVs were optimised by repeated cross-validation by varying the LVs from 0 to 15, the optimal complexity is then defined using the RMSECV. In the present work, the algorithm presented in (Roger et al., 2020a) was implemented in MATLAB 2017b, Natick, USA.

3. Results

The spectra of different fruits and olive oil samples are presented in Fig. 2. In the case of apples and grapes (Fig. 2A and B), the spectra in the range of 499 – 670 nm are related to the pigment composition of the skin of the fruits. Furthermore, the range of 670 – 1018 nm corresponds to the 3rd overtones of C—H and O—H bonds and is widely used in analysis for fruit products to predict Brix and DM content. In the case of olives, the spectra were in the range 669 – 1122 nm, which is related to the 3rd overtones of the C—H and O—H bonds (Osborne, 2006). In the case of apricot and olive oils (Fig. 2C, E and F), the spectra correspond to the 1st and the 2nd overtones as the spectral range was > 1000 nm. In the spectra of the fruits (apples, grapes, apricots and olives), a difference in the global spectrum intensities can be observed. Such a difference is a clear indication of scattering effects (additive and multiplicative). In the case of olive oils, there is no such visible global intensity differences, however, as the spectra were acquired in transmission there could be scattering effects in the spectra.

3.1. Spectra

The spectra from all the samples were pre-processed with 2nd derivative and mean centered. This was done as a standard method for comparison with the SPORT pre-processing approach. Separate PLSR models were developed for each material and the results are presented in Fig. 3. For PLSR modeling, the spectra were distributed in a ratio of 70/30 % as the calibration and test set utilizing the Kennard-Stone (KS) algorithm. The number of latent variables (LVs) was optimized using the 10-fold 'venetian blind' cross validation approach. In Fig. 3, the blue points represent the calibration set and the red points represents the test set. In the case of olive oils, with calibration to predict LOL, the PLSR modeling resulted in the selection of just 1 L V indicating that the 2nd derivative operation was able to resolve the peaks corresponding to the LOL triglycerides. In the case of the fruits, the number of LVs ranged from 4 to 7. The R_p^2 was greater than 0.90 in the case of the grape and olive dataset. In the case of apple, olive fruit and apricot, the R_p^2 range was 0.69 – 0.81.

The SNV + 2nd derivative showed an improvement in model performance compared to just 2nd derivative as presented in Fig. 4. For the apple dataset, R_{pred}^2 increased by 2 % and RMSEP decreased by 2.8 %. For the grape dataset, R_{pred}^2 was increased by 3 % and the RMSEP decreased by 23 %. To olive oils dataset, R_{pred}^2 increased by 2 % and the RMSEP decreased by 21 %. For olive fruit dataset, R_{pred}^2 increased by 7 % and the RMSEP decreased by 19 %. For the apricot dataset, no change in R_{pred}^2 and RMSEP was observed.

3.2. Modelling with SPORT pre-processing

Fig. 5 shows the modeling results from the SPORT pre-processing fusion for the same datasets presented in Figs. 3 and 4. A summary of improvements in model performance is presented in Table 2. NIRS prediction modeling of ingredients (soluble solids, organic acids, dry matter, fatty acids) with the SPORT approach for all the food materials resulted in an increase in R_{pred}^2 and a decrease in RMSEP.

In comparison to the SNV + 2nd derivative, the improvement with the SPORT pre-processing approach was limited to the samples measured in reflection mode i.e. the apple, olive fruit and apricot datasets. In the case of transmission measurements i.e. grapes and olive oils, the improvements were limited. This was because in the presented use of the SPORT pre-processing, the aim was to eliminate the scatter from the data which in the case of transmission is low compared to reflection. However, the principle presented regarding optimal fusion of pre-

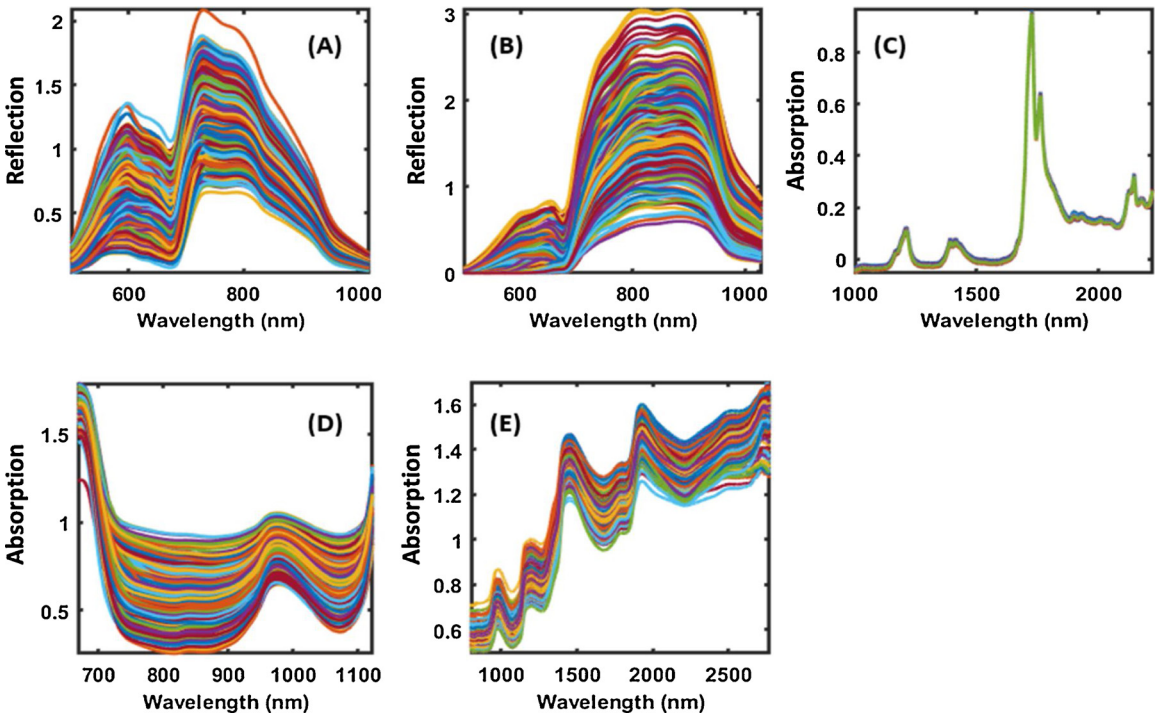

Fig. 2. All individual spectra from the six different datasets. (A). Apple, (B). Grape, (C). Olive oil, (D). Olive fruit, and (E). Apricot.

treatments remains valid if the pre-treatment methods used correspond to the factors that need to be removed. In summary, pre-processing fusion of scatter correction methods improves reflection spectra more than transmission spectra as the scattering is far greater in the reflection mode.

A summary of all the models and prediction error is presented in Table 3. The results showed that for the case of reflection the SNV + 2nd derivative performed better than only 2nd derivative and the SPORT approach performed better than SNV + 2nd derivative as well as 2nd derivative alone. The improvement in model performance was obtained because the SPORT approach used information from multiple pre-processing techniques (Table 4). This was possible through the sequential extraction of latent variables from different pre-processing treatments considered as different blocks. In the case of the apple dataset, the SPORT approach used a fusion of MSC and VSN pre-processing. In the case of grapes, a fusion of raw data with VSN and 2nd derivative

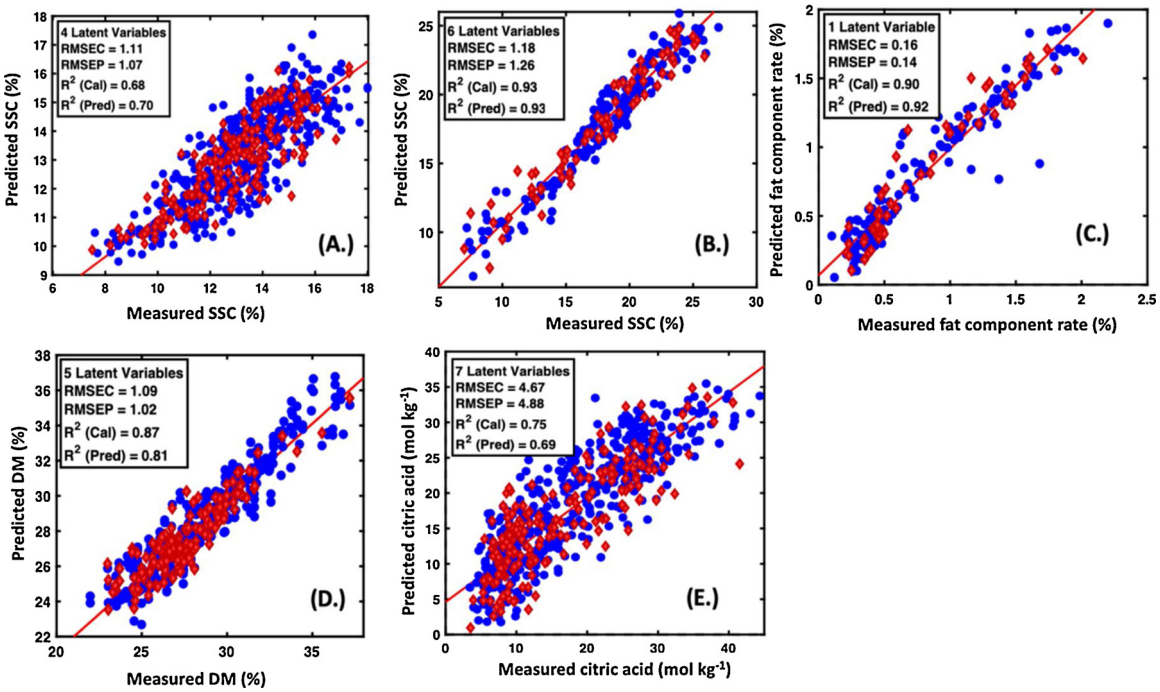

Fig. 3. PLS models based on 2nd derivative spectra. (A). Brix prediction in apples, (B). Brix prediction in grapes, (C). linoleic acid prediction in olive oils, (D). dry matter prediction in olive fruits, and (E). citric acid prediction in Apricot. The blue dots explains the calibration set (70 %) and the red dots explains the test set (30 %) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

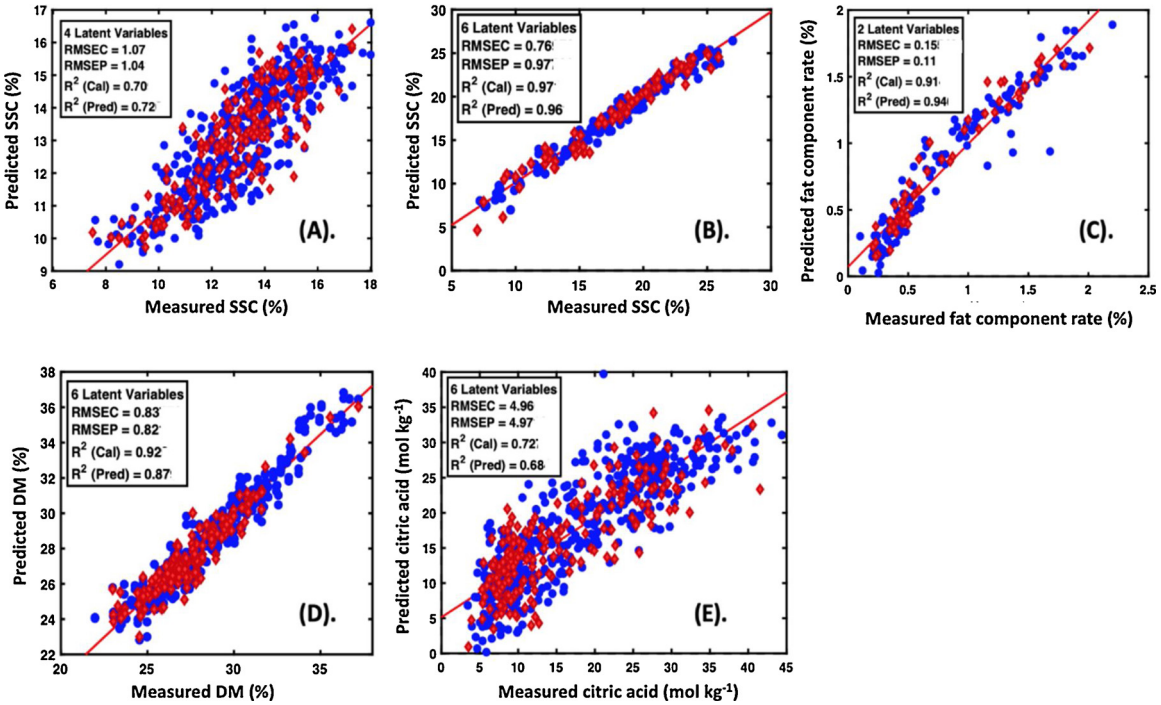

Fig. 4. PLS models based on standard normal variate + 2nd derivative spectra. (A). Brix prediction in apples, (B). Brix prediction in grapes, (C). Linoleic acid prediction in olive oils, (D). dry matter prediction in olive fruits, and (E). citric acid prediction in Apricot. The blue dots explains the calibration set (70 %) and the red dots explains the test set (30 %) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

preprocessing was performed. In the case of olive oils, a fusion of VSN, SNV and 2nd derivative was identified. In the case of olive fruits, a combination of VSN and 2nd derivative was optimal. In comparison to the primary work related to olive fruits, the combination of 2nd derivative and VSN was found optimal by (Sun, Subedi et al. 2020). However, 2nd derivative and VSN were used one after another and not in a data fusion manner as processed by SPORT. The SPORT approach had a much lower RMSEP (0.75) compared to the primary work (0.90) by (Sun, Subedi et al. 2020). In the case of apricot, a fusion of MSC, SNV and 2nd derivative led to an increased model performance.

To have a better understanding of what complementary information from selected pre-processing techniques are being added, the regression

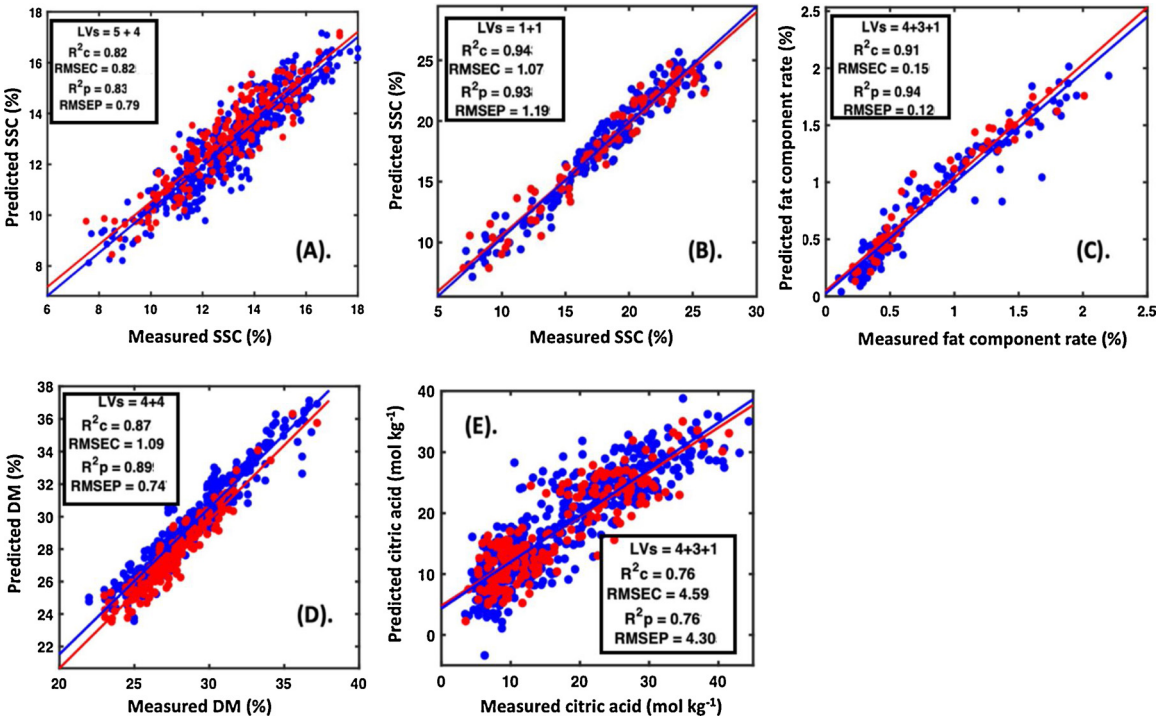

Fig. 5. SPORT preprocessed models. (A). Brix prediction in apples, (B). Brix prediction in grapes, (C). Linoleic acid fat prediction in olive oils, (D). dry matter prediction in olive fruits, and (E). citric acid prediction in Apricot. The blue dots explains the calibration set (70 %) and the red dots explains the test set (30 %) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Table 2
Summary of improvement with the SPORT pre-processing.

Dataset	Measurement mode	Improvement with SPORT compared to 2nd derivative		Improvement with SPORT compared to SNV + 2nd derivative	
		Increase in R^2 (%)	Decrease in RMSEP (%)	Increase in R^2 (%)	Decrease in RMSEP (%)
Apple	Reflection	20	25	16	23
Grapes	Transmission	1	4	N.I.	N.I.
Olive oils	Transmission	2	21	N.I.	N.I.
Olive fruit	Reflection	11	26	3	8
Apricot	Reflection	10	11	11	13

Table 3
A summary of models obtained after different pre-processing approaches. Chemical units can be referred in Table 1. Abbreviations are: Latent variables (LVs), root mean squared error of prediction (RMSEP), standard normal variate (SNV).

Dataset	2nd derivative		SNV + 2nd derivative		SPORT	
	LVs	RMSEP	LVs	RMSEP	LVs	RMSEP
Apple	4	1.07	4	1.04	5 + 4	0.8
Grapes	6	1.26	6	0.97	1 + 1	1.2
Olive oils	1	0.14	2	0.12	1 + 1 + 1	0.12
Olive fruit	5	1.02	6	0.82	4 + 4	0.75
Apricot	7	4.88	6	4.98	4 + 3 + 1	4.30

Table 4
Optimal number latent variables selected from each block for the respective datasets. Abbreviations are: multiplicative scatter correction (MSC), variable sorting for normalization (VSN) and standard normal variate (SNV).

Datasets/Pre-processing	Raw data	MSC	VSN	SNV	2nd derivative
Apple		5	4		
Grapes	1		1		5
Olive oil			1	1	1
Olive fruits			4		4
Apricot		4		3	1

vectors from each bock are presented in Fig. 6–10. To facilitate the comparison the regression vectors are plotted in arbitrary scales. In the case of the apple dataset, the SPORT selected the VSN and the MSC as optimal pre-processing techniques. The VSN pre-processing highlighted extra wavelengths corresponding to CH_2 and CH bonds which were missed during the MSC pre-processing (Fig. 6). CH_2 and CH bonds could be linked to the sugar components in the fruits which are linked with the Brix (Osborne, 2006).

In the case of the grapes dataset, the SPORT identified raw, VSN and 2nd derivative as optimal pre-processing techniques. Both the MSC and the VSN (Fig. 7) were dominated by the color information whereas the 2nd derivative captured the CH2, ArOH and H_2O wavelength regions (Osborne, 2006).

In the case of the olive oil dataset, the SPORT identified the VSN, SNV and 2nd derivative as optimal pre-processing techniques. Information around 1700 nm (Fig. 8) was captured by all the three pre-processing techniques and can be explained as the first overtones of the CH_3, CH_2 and CH bonds in the triglycerides (Osborne, 2006). However, SNV captured some extra information around 2200 nm which also corresponds to the CH_3 combination band region (Osborne, 2006). The complementary role was the fusion of information of the first overtones and the combination bonds.

In the case of the olive fruits dataset, the SPORT identified the VSN and the 2nd derivative. The 2nd derivative captured the information about the OH bond whereas the VSN extracted information mainly about ArOH, meaning that apart from moisture there is also some ArOH compound that are correlated with the final dry matter of fruits (Fig. 9).

In the case of citric acid there were a number of peaks present in the regression vector from the different pre-processing (Fig. 10). In case of SNV, there are peaks at 1000 nm and 1700 nm which were not captured by the MSC and the 2nd derivative. The 2nd derivative captured something distinct at 1900 nm.

4. Discussion

In the present study, the potential of the SPORT-based pre-processing fusion to improve NIRS prediction models is presented for several food materials. Four different scatter correction techniques (SNV, MSC, VSN and 2nd derivative) were used for the fusion. NIRS data often suffers from scattering effects and these effects, if not properly modeled, can affect the performance of the models. In the case of fruit, the problem of scattering is widely known (Lu et al., 2020), and often multiple combinations of pre-processing techniques are deployed (Nicolai et al., 2007; Saeys et al., 2019). Fusion of multiple scatter correction techniques is still unexplored, even though no single pre-processing is perfect. Different scattering techniques have advantages as well as disadvantages (Roger et al., 2020b).

Different pre-processing techniques can provide complementary information which can be used in a sequential approach to improving models (Roger et al., 2020a). In the present work, the complementary fusion of differently pre-processed data with the SPORT approach showed model improvements for all the food materials. In the case of apple dataset, the MSC and the VSN were able to capture complementary peaks related to the sugar components. In the cases of grape dataset, the 2nd derivative revealed the main peaks related to sugar

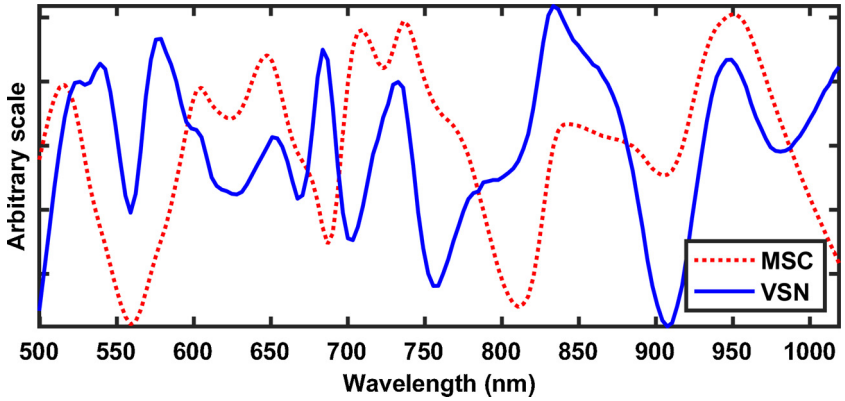

Fig. 6. Regression vectors from multiplicative scatter correction (MSC) (dashed red) and variable sorting for normalization (VSN) (solid blue) pre-processed blocks of the apple dataset (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

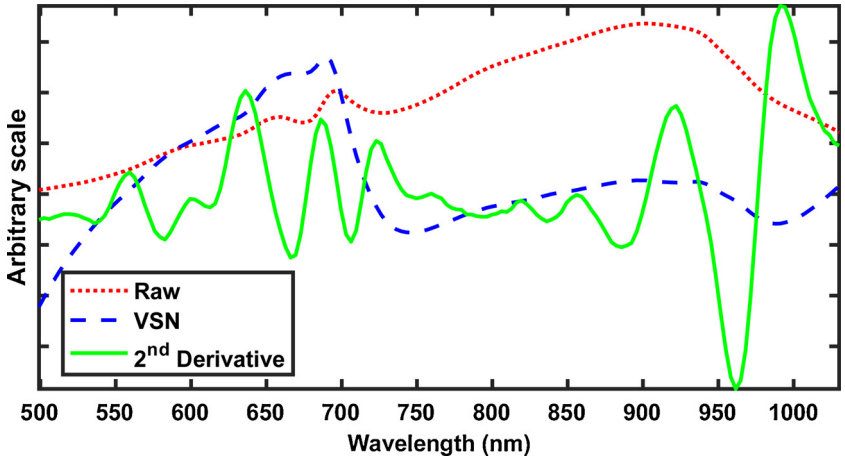

Fig. 7. Regression vector from Raw (dotted red), variable sorting for normalization (VSN) (dashed blue) and 2nd derivative (solid green) pre-processed blocks of the grape dataset (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

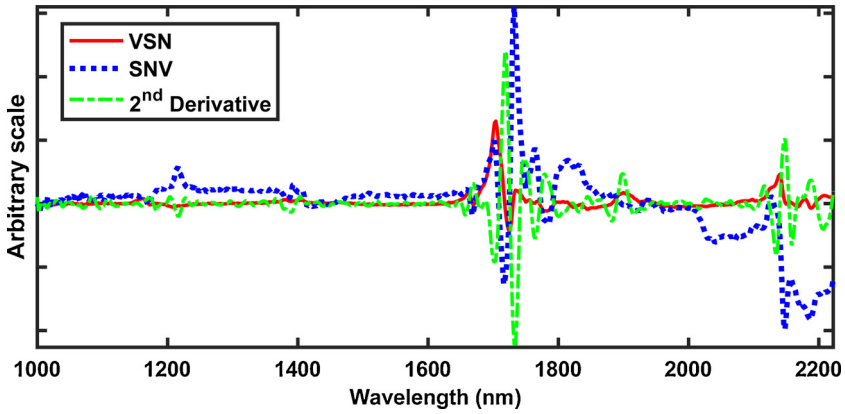

Fig. 8. Regression vector from variable sorting for normalization (VSN) (solid red), standard normal variate (SNV) (dotted blue) and 2nd derivative (solid green) pre-processed blocks of the olive oil dataset (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

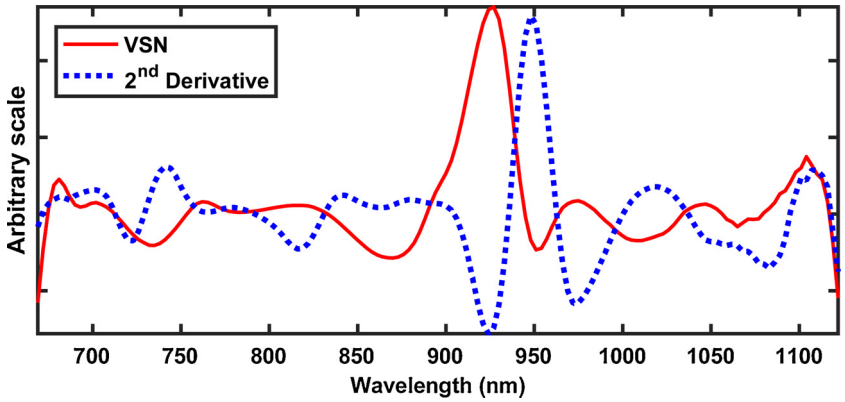

Fig. 9. Regression vector from variable sorting for normalization (VSN) (solid red) and 2nd derivative (dotted blue) pre-processed blocks of the olive fruit dataset (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

components whereas the VSN was mainly limited to the color information. In the case of the olive oil dataset, the SNV was able to capture also the combination bonds information in the regression vector compared to other pre-processing. In the case of the olive fruit dataset, the VSN and the 2nd derivative captured different peaks, highlighting that different pre-processing algorithms are not capturing the same information and rather capturing different things. In the case of apricot as well there are multiple distinct peaks captured by different pre-processing algorithms. A point to be understood is that in the case of NIRS, due to the highly overlapping peaks, it is difficult to clearly assign the peaks to any chemical component. It could also be possible that the peaks captured by different pre-processing's have a secondary indirect correlation with the main property to be predicted. However, in all the cases presented in this work, the fusion of the pre-processing methods resulted in improved model accuracy.

The improvements in the models were much greater in the case of reflection measurement mode on fruit compared to the transmission mode on grapes and oils samples. There could be because the reflection mode is more prone to scattering effects than the transmission mode. The reflection is performed in a non-contact way and is infamous to

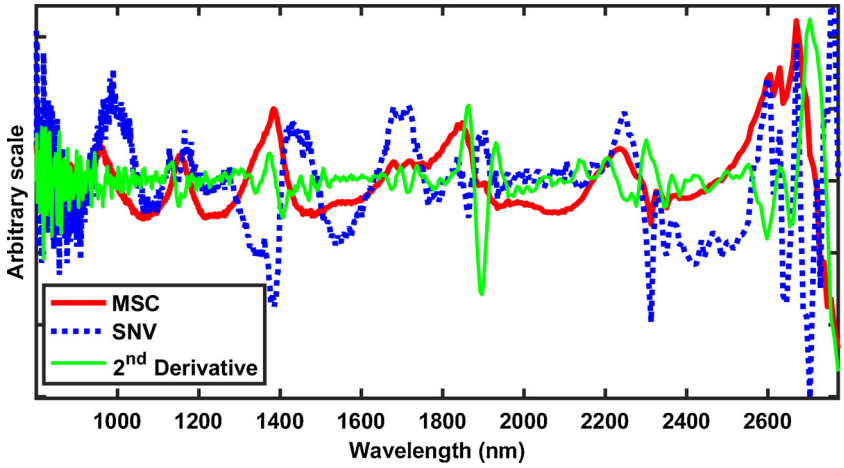

Fig. 10. Regression vector from, multiplicative scatter correction (MSC) (solid red), standard normal variate (SNV) (dotted blue) and 2nd derivative (solid green) pre-processed blocks of the apricot dataset (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

have huge scattering (in the case of fruits) accompanied due to interaction of the light and the multiple internal reflections when it penetrates the surface of material before being reflected. Further, the fruit peel and flesh attenuate these effects. In the present work, fusion of different scatter correction algorithms improved the model performance for the samples measured in reflection mode. There was little improvement in the case of transmission datasets of grapes and olive oils as, apparently, scattering is less of a problem in these samples. For NIR reflection measurements of fruit quality aspects, optimal fusion of pre-processing algorithms with SPORT pre-processing may be recommended to improve prediction power.

5. Conclusions

Different pre-processing algorithms can add complementary information to the model leading to increased prediction accuracy. In the case of spectroscopy, fusion of multiple pre-processing methods previously was limited to ensemble approaches. In the present work, a multi-block data fusion inspired pre-processing fusion approach called SPORT showed improvements in the NIRS predictive model performance for a range of agri-food materials. In the cases of fruit measured in reflection mode, the improvement led to as much as a 20 % increase in R^2_{pred} and a 25 % decrease in RMSEP compared to the standard 2nd derivative pre-processing. The improvement for olive oil and grape berries measured in transmission mode the improvements were limited. The SPORT approach require the user to define the pre-processing order. The pre-processing order will not affect the model performance. An easy way to define the order is by the computational cost of the pre-processing techniques with SPORT starting from pre-processing's with low cost at first and swiftly moving to more complex pre-processing techniques. The benefit of the SPORT approach is that it is an automated technique to select and perform fusion of multiple pre-processing techniques.

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Data curation, Investigation. **Jean Michel Roger:** Formal analysis, Software, Visualization. **Douglas N. Rutledge:** Formal analysis, Methodology, Software. **Ernst Woltering:** Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgment

Prof. Kerry B. Walsh and Dr. Phul Subedi from Central Queensland University, Australia for sharing the Olive fruit dataset used in the study.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.postharvbio.2020.111271.

References

Arendse, E., Fawole, O.A., Magwaza, L.S., Opara, U.L., 2018. Non-destructive prediction of internal and external quality attributes of fruit with thick rind: a review. J. Food Eng. 217, 11–23.
Armenta, S., Moros, J., Garrigues, S., Guardia, M.D.L., 2010. The use of near-infrared spectrometry in the olive oil industry. Crit. Rev. Food Sci. Nutr. 50 (6), 567–582.
Awhangbo, L., Bendoula, R., Roger, J.M., Béline, F., 2020. Multi-block SO-PLS approach based on infrared spectroscopy for anaerobic digestion process monitoring. Chemom. Intell. Lab. Syst. 196, 103905.
Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43 (5), 772–777.
Bi, Y., Yuan, K., Xiao, W., Wu, J., Shi, C., Xia, J., Chu, G., Zhang, G., Zhou, G., 2016. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. Anal. Chim. Acta 909, 30–40.
Bian, X., Wang, K., Tan, E., Diwu, P., Zhang, F., Guo, Y., 2020. A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples. Chemom. Intell. Lab. Syst. 197, 103916.
Biancolillo, A., Liland, K.H., Måge, I., Næs, T., Bro, R., 2016. Variable selection in multi-block regression. Chemom. Intell. Lab. Syst. 156, 89–101.
Biancolillo, A., Næs, T., Bro, R., Måge, I., 2017. Extension of SO-PLS to multi-way arrays: SO-N-PLS. Chemom. Intell. Lab. Syst. 164, 113–126.
Biancolillo, A., Næs, T., Cocchi, M., 2019. Chapter 6 - the sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions. Data Handling in Science and Technology 31. Elsevier, pp. 157–177.
Bureau, S., Ruiz, D., Reich, M., Gouble, B., Bertrand, D., Audergon, J.-M., Renard, C.M.G.C., 2009. Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit. Food Chem. 115 (3), 1133–1140.
Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal. Chem. 78 (13), 4281–4290.
Engel, J., Gerretzen, J., Szymańska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L.M.C., 2013. Breaking with trends in pre-processing? TrAC Trends Anal. Chem. 50, 96–106.
Firmani, P., Nardecchia, A., Nocente, F., Gazza, L., Marini, F., Biancolillo, A., 2020. Multi-block classification of Italian semolina based on Near Infrared Spectroscopy (NIR) analysis and alveographic indices. Food Chem. 309, 125677.
Galtier, O., Dupuy, N., Le Dréau, Y., Ollivier, D., Pinatel, C., Kister, J., Artaud, J., 2007. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. Anal. Chim. Acta 595 (1), 136–144.
Gerretzen, J., Szymańska, E., Jansen, J.J., Bart, J., van Manen, H.-J., van den Heuvel, E.R., Buydens, L.M.C., 2015. Simple and effective way for data preprocessing

selection based on design of experiments. Anal. Chem. 87 (24), 12096–12103.
Gómez-Caravaca, A.M., Maggio, R.M., Cerretani, L., 2016. Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review. Anal. Chim. Acta 913, 1–21.
Guo, Q., Wu, W., Massart, D.L., 1999. The robust normal variate transform for pattern recognition with near-infrared data. Anal. Chim. Acta 382 (1), 87–103.
Isaksson, T., Næs, T., 1988. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Appl. Spectrosc. 42 (7), 1273–1284.
Kamruzzaman, M., Makino, Y., Oshita, S., 2015. Non-invasive analytical technology for the detection of contamination, adulteration, and authenticity of meat, poultry, and fish: a review. Anal. Chim. Acta 853, 19–29.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11 (1), 137–148.
Lammertyn, J., Peirs, A., De Baerdemaeker, J., Nicolaï, B., 2000. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biol. Technol. 18 (2), 121–132.
Lu, R., Van Beers, R., Saeys, W., Li, C., Cen, H., 2020. Measurement of optical properties of fruits and vegetables: a review. Postharvest Biol. Technol. 159, 111003.
Måge, I., Smilde, A.K., van der Kloet, F.M., 2019. Performance of methods that separate common and distinct variation in multiple data blocks. J. Chemom. 33 (1), e3085.
Martens, H., Nielsen, J.P., Engelsen, S.B., 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Anal. Chem. 75 (3), 394–404.
Nicolai, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. Postharvest Biol. Technol. 46 (2), 99–118.
Niimi, J., Tomic, O., Næs, T., Jeffery, D.W., Bastian, S.E.P., Boss, P.K., 2018. Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition. Food Chem. 256, 195–202.
Osborne, B.G., 2006. Near-Infrared Spectroscopy in Food Analysis. Encyclopedia of Analytical Chemistry.
Pasquini, C., 2018. Near infrared spectroscopy: a mature analytical technique with new perspectives – a review. Anal. Chim. Acta 1026, 8–36.
Rabatel, G., Marini, F., Walczak, B., Roger, J.-M., 2020. VSN: variable sorting for normalization. J. Chemom. 34 (2), e3164.
Rinnan, Å., Berg, Fvd., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28 (10), 1201–1222.
Roger, J.-M., Chauchard, F., Bellon-Maurel, V., 2003. EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. Chemom. Intell. Lab. Syst. 66 (2), 191–204.
Roger, J.-M., Biancolillo, A., Marini, F., 2020a. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. Chemom. Intell. Lab. Syst. 199, 103975.
Roger, J.-M., Boulet, J.-C., Zeaiter, M., Rutledge, D.N., 2020b. Pre-processing methods★. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Elsevier.
Saeys, W., Do Trong, N.N., Van Beers, R., Nicolai, B.M., 2019. Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review. Postharvest Biol. Technol. 158.
Sierra, V., Aldai, N., Castro, P., Osoro, K., Coto-Montes, A., Oliván, M., 2008. Prediction of the fatty acid composition of beef by near infrared transmittance spectroscopy. Meat Sci. 78 (3), 248–255.
Smilde, A.K., Måge, I., Næs, T., Hankemeier, T., Lips, M.A., Kiers, H.A.L., Acar, E., Bro, R., 2017. Common and distinct components in data fusion. J. Chemom. 31 (7), e2900.
Song, Y., Westerhuis, J.A., Smilde, A.K., 2020. Separating common (global and local) and distinct variation in multiple mixed types data sets. J. Chemom. 34 (1), e3197.
Sun, X., Subedi, P., Walker, R., Walsh, K.B., 2020. NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pretreatment. Postharvest Biol. Technol. 163, 111140.
Walsh, K.B., McGlone, V.A., Han, D.H., 2020. The uses of near infra-red spectroscopy in postharvest decision support: a review. Postharvest Biol. Technol. 163, 111139.
Wang, H.L., Peng, J.Y., Xie, C.Q., Bao, Y.D., He, Y., 2015. Fruit quality evaluation using spectroscopy technology: a review. Sensors 15 (5), 11889–11927.
Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., Yu, R.-Q., 2008. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. Anal. Chim. Acta 616 (2), 138–143.