Postharvest Biology and Technology 171 (2021) 111348

Contents lists available at ScienceDirect

Postharvest Biology and Technology

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

# Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach

Puneet Mishra [a,*], Ernst Woltering [a,b], Bastiaan Brouwer [a], Esther Hogeveen-van Echtelt [a]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] Horticulture and Product Physiology Group, Wageningen University, Droevendaalsesteeg 1, P.O. Box 630, 6700AP, Wageningen, the Netherlands

ARTICLE INFO

*Keywords:*
Interval partial least-squares regression
Covariate selection
Chemometric
Non-destructive
Fruit-quality

ABSTRACT

To obtain robust near-infrared (NIR) spectroscopy data calibration models, variable selection and model updating with recalibration approaches were used for predicting quality parameters in pear fruit. For variables selection, interval partial least-squares regression and covariate selection approaches were used and compared. Model updating with recalibration was performed by incorporating a few new samples in the calibration set of existing batch data. The interaction of variable selection and model updating was also explored. The results showed that with variable selection, the model performance when tested on a new independent batch of fruit was greatly improved. Further, the model updating with only a few new samples resulted in a reduction of the bias when tested on the new batch. In the case of MC prediction, the variable selection reduced the bias from 1.31 % to 0.19 % and the RMSEP from 1.44 % to 0.58 %, compared to the standard partial least-squares regression (PLS2R). In the case of SSC prediction, the variable selection reduced the bias from -0.62 % to 0.07 % and the RMSEP from 0.90 % to 0.63 %, compared to the standard PLS2R. With a combination of variable selection and model updating the bias and RMSEP were further reduced. The interval-based method performed better compared to the filter-based method. As few as only 10 samples from the new batch already lead to a significant improvement in model performance. In the case of MC, spectral regions of 749-759 nm and 879-939 nm were identified as the most important region. In the case of the SSC, 709-759 nm and 789-999 nm were found to be important spectral regions. Robust models made on selected variables combined with model updating strategy can support to make NIR spectroscopy a preferred choice for non-destructive assessment of quality features of fresh fruit.

## 1. Introduction

Fresh fruit are widely traded across the world. To make long-distance transport possible, fruit are often harvested in an immature stage. Fruit harvest date is often decided based on parameters such as soluble solids content (SSC) and moisture content (MC). This is because SSC and MC contribute to indirect estimations of fruit maturity and quality, where low SSC and low MC values implicate unripe, less tasty fruit (Palmer et al., 2010; Travers et al., 2014). A common non-destructive tool to achieve this is with near-infrared (NIR) spectroscopy (Nicolai et al., 2007; Wang et al., 2015). In NIR spectroscopy, spectra of the fruit are acquired using dedicated spectrometers and calibration models are used to provide output as a prediction of quality parameters (Lu et al., 2020). Further, the prediction is combined with background physiological knowledge, such as a range of MC and SSC for raw fruit, to make decisions. Apart from harvest decisions, NIR spectroscopy is increasingly being integrated with fruit sorting lines, ripening monitoring and for decisions on maturity levels of fruit in storage (Walsh et al., 2020).

NIR spectroscopy data is multivariate and made up of several underlying peaks related to multiple chemical compounds such as water, sugar, protein and fats (Lin and Ying, 2009). The modelling performed on the NIR spectroscopy data involves identification of the underlying peaks to avoid the collinearity problem and later using the information extracted from these corresponding peaks to calibrate the model (Saeys et al., 2019). Based on the requirement, classification or regression analysis can be performed. A common technique used for NIR

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

https://doi.org/10.1016/j.postharvbio.2020.111348
Received 10 July 2020; Received in revised form 31 August 2020; Accepted 2 September 2020
Available online 25 September 2020
0925-5214/© 2020 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

spectroscopy data modelling is the partial least-squares regression (PLS2R) (Wold et al., 2001). Here, the 2 in the PLSR indicates that multi-responses are considered by the same PLSR model. PLS2R works by identifying the latent variables (LVs) which explains the variance in the response variables. To perform the regression, the NIR spectroscopy data is transformed by projecting it to the identified LVs followed by a multi-linear regression. However, models developed on NIR spectroscopy data usually work well for a single fruit batch but fail to perform well when tested on a different batch (Teh et al., 2020). Often when tested on data from a different batch, high bias and error are prevalent. Several physical, chemical and environmental effects can be accounted for the failure of models (Zeaiter et al., 2006).

A reason for the poor performance of a model when applied to a new fruit batch could be due to the existing model being sub-optimal. Sub-optimal modelling in NIR spectroscopy could results as NIR data is a mixture of several overlapping peaks which are sometimes difficult to extract with methods such as PLS2R (Nørgaard et al., 2000). However, advanced variable selection methods were found to be complementing the PLS2R modelling to further optimise PLS-based models (Mehmood et al., 2012; Mehmood et al., 2020). Variable selection often results in the removal of variables that were estimated as being of least importance by either identifying and keeping the important wavelength intervals over the spectral range or filtering discrete wavelengths. By doing so, the burden over the PLS2R to identify the optimal LVs is reduced and thus results in more efficient modelling (Mehmood et al., 2012; Mehmood et al., 2020). NIR spectra (750 −2500 nm) are made up of multiple peaks with specific wavelengths corresponding to chemical components (Mishra et al., 2017), such as for water and sugar prediction, the 3$^{rd}$ overtones of OH and CH bonds can be related to the spectral range of 750-950 nm. This means that making a model for predicting e.g. MC and SSC in the fruit, using the information from wavelengths outside the region of OH and CH overtones may lead to sub-optimal models.

In the domain of chemometrics, there are two main types of variable selection techniques exist i.e., interval-based and the filter-based techniques (Mehmood et al., 2012; Mehmood et al., 2020). The interval-based techniques select a sub-region over the signal which are the most predictive of the response variables. The interval-based techniques are useful when the data has continuous variables such as in the case of the NIR spectroscopy (Nørgaard et al., 2000). The filter-based techniques aim to filter out the individual variables based on some criteria such as the maximum covariance or with the use of a user-defined threshold (Roger et al., 2011). Particular to the NIR spectroscopy of fresh fruit, variable selection with interval-based approaches have shown improved and generalised model performances. For examples, model based on selected intervals improved the firmness prediction in mango (Valente et al., 2009), glucose and sucrose content prediction in potatoes (Rady and Guyer, 2015), and total soluble solids (TSS), dry matter (DM), flesh colour and acidity prediction in mango fruit (Nordey et al., 2017).

One of the ways to deal with the poor performance of models when applied to a new batch is with the model updating. A study related to SSC prediction in apple fruit suggested that including more variability in the calibration set as new samples from varying orchards, season and cultivars, improved the model robustness (Peirs et al., 2003). Similarly, in another study related to SSC prediction in apple fruit, incorporating extra variability in the calibration set improved the predictive performance of apple models for different cultivar, season, shelf-life and origin of the fruit (Bobelyn et al., 2010). A study related to plum fruit showed that incorporating extra samples from multiple cultivars improved the generalisability of NIR models for plum quality parameters prediction (Louw and Theron, 2010). In the case of mango fruit, incorporating data from multiple seasons improved the robustness of NIR models compared to models made on a single season data (Rungpichayapicheta et al., 2016). The model updating with recalibration is performed by adding some extra measurements i.e., NIR spectroscopy and reference, from the new batch to the existing data or combining data from multi-season (Rungpichayapicheta et al., 2016), multi-cultivars (Louw and Theron, 2010) and several measurement conditions (Peirs et al., 2003; Bobelyn et al., 2010; Nordey et al., 2017). The old model is then recalibrated with the new samples and used for the prediction of the new batch. Often, the model updating results in an improvement in model performance in terms of reduction in bias and error. However, the main drawback of the model updating with recalibration is that it requires new samples which are not always available. Also, it is not clear how many new samples are required for the model updating. In developing robust models, the first aim should be to optimise the PLS2R models (Nascimento et al., 2016) such that they can be used with acceptable performance on different batches without the need of the model updating and extra measurements. Secondly, if the optimised models are still poor in performance then model updating should be incorporated to enhance the models. Both the model optimisation and model updating with recalibration require extra efforts, but the model optimisation step does not require any new measurements.

The overall aim of this study was to achieve robust NIR spectroscopy models which perform well when tested on different fruit batches. To attain that, two variable selection approaches i.e., iPLS2R and covariate selection (CovSel), were compared. The models were explored for MC and SSC prediction in individual pear fruit. Further, a combination of model updating with recalibration and variable selection was explored to identify a minimum number of samples required to perform model updating.

## 2. Material and methods

### 2.1. Plant material

Two batches of 'Conference' pear fruit (*Pyrus communis L.*) were measured. 239 samples from Batch 1 and 240 samples from Batch 2. Both batches contained a mix of fruit from 10 different orchards throughout The Netherlands i.e., Randwijk, Broex, Tiel, Biolet, Deil, Zeeland, Westwoud, near Utrecht, west of Utrecht and near Rotterdam, and 1 from Belgium i.e., Sint Truiden, with one orchard from The Netherlands i.e., Randwijk, delivering two groups: one with normal irrigation and one subjected to an average soil water tension of −100 kPa during the month prior to harvest. The soil type was either the river clay or the sea clay. The samples were received at Wageningen, The Netherlands after 1-2 days of harvest in the middle of harvest season of year 2019. After harvest pear fruit were either stored under regular controlled atmosphere conditions for pear fruit (0.7 % $CO_2$ and 3 % $O_2$ at -0.5 °C and > 95 % RH) for 8 months or analysed immediately. The difference between two batches was that the batches were measured 8 months apart. To obtain generalised models, the orchards had been selected for a wide diversity of pear fruit, based on size, amount of skin rusting and shape.

### 2.2. Visible and near-infrared spectroscopy measurements

The spectral measurements were carried out with a portable spectrometer (Felix F-750, Camas, WA, USA). The Felix utilises a Carl Zeiss MMS-1 spectrometer to record the reflected light in the spectral range of 310-1135 nm with a spectral resolution of 8-13 nm. The spectrometer utilizes a Xenon Tungsten Lamp for illumination and a built-in white painted reference standard for estimating the reflectance. The data acquisition was performed by placing the fruit at the sample holder and by manually pressing the scan button on the Felix device. For a single pear fruit, spectral measurements were performed at centre belly part. The final scan was an automatic average of 6 scans from the same spot. The samples for reference measurements were taken to include at least part of the area recorded by the spectral measurements. The data were automatically radiometric calibrated by the Felix device and the reflectance spectra were extracted as excel files using the "Data-Viewer" software (Felix Instruments, Camas, WA, USA). The radiometric calibration was performed as per Eq. 1:

$$Reflectance = \frac{S - A - D}{W - D} \quad (1)$$

where $S$ is fruit spectra (acquired with shutter open, lamp on), $A$ is ambient light spectra (acquired with shutter open, lamp off), $D$ is dark reference spectra (acquired with shutter closed, lamp off), and $W$ is white reference spectra (acquired with shutter closed, lamp off).

### 2.3. Reference measurements

MC and SSC measurements were performed as the reference to correlate with NIR spectroscopy data. A schematic of the sampling procedure is shown in Fig. 1. After NIR spectroscopy measurements on the fruit samples, a 1 cm thick slice was cut from the equator of the pear fruit belly, which was subsequently divided into 4 equal parts. Two of these parts were used to determine MC and SSC. MC was determined using an electronic balance XS10001 L (Mettler-Toledo GmbH, Giessen, Germany) by recording the weight of the parts before and after drying in a hot-air oven (FP 720, Binder GmbH, Tuttlingen, Germany) at 80 °C for 96 h. SSC of extracted pear fruit juice was determined using a handheld refractometer (HI 96801, Hanna Instruments Inc, Woonsocket, RI, USA).

### 2.4. Data analysis

#### 2.4.1. Spectral pre-processing

The spectral range was reduced from 310-1135 nm to the NIR range 700-1135 nm. This was done for two reasons: first to remove the influence of fruit colour from the models and second to focus on the 3rd overtones related to O-H and C-H bonds present in the spectral range of 700-1135 nm. The MC and SSC can be correlated with the O-H and C-H containing compounds. The spectra were smoothened with Savitzky-Golay (SavGol) smoothing utilising a default window size of 15 and fitting a second order polynomial. The scatter was corrected using the standard normal variate (SNV) transform (Barnes et al., 1989). Further to reveal the underlying peaks, the 2nd derivative was estimated utilising a default window size of 15 and fitting a second order polynomial. In all the cases, the models were developed on Batch 1 and tested on Batch 2.

#### 2.4.2. Partial least-squares regression

PLS2R is a common chemometric technique used for NIR spectroscopy data modelling (Saeys et al., 2019; Wold et al., 2001). PLS2R deals with the multi-co-linearity in the multivariate signal by extracting the underlying peaks as the LVs. The LVs were extracted having maximum covariance with the response variables. To perform simultaneous prediction of MC and SSC, a multi-response PLS2R known as PLS2R was used. The PLS2R is different from PLS2R in the way that it also decomposes the response matrix into scores and loading, and the covariance with the predictor block is maximised using the scores. Further, in the case of prediction, the scores are predicted for the response variables, which are multiplied with the loading to obtain actual responses. Before feeding to PLS2R, the data were mean centred. The LVs were selected utilising Venetian-blind cross-validation (10 random blocks) and the output of the regression is presented as coefficient of determination and root mean squared errors (RMSE). The corresponding MC and SSC measurements were also mean centred before regression analysis. Outlying samples were identified and kept out of modelling utilizing the PLS inner relation plots.

#### 2.4.3. Interval partial least-squares regression

NIR spectroscopy data consists of a high number of variables. Preselecting the important wavelengths makes the modelling task easier and provides a better explanation for the modelling relation and the variables (Mehmood et al., 2012; Mehmood et al., 2020). In the present work, a common chemometric algorithm called interval partial least-squares regression (iPLS2R) was used for variable selection (Nørgaard et al., 2000; Zou et al., 2010). An interval is a subset of continuous wavelengths. iPLS2R selects intervals in following steps:

1 An interval size (**n**) is defined by the user.
2 For a spectra set of dimension **p** × **m**, where **p** is the sample size and **m** are the wavelengths (nm), the spectra are divided into **m/n** intervals.
3 Using each spectral interval separate PLS2R models are developed and optimized using Venetian-blind cross-validation (10 random blocks).
4 The cross-validation error obtained for each interval is compared to the PLS2R model developed using the full spectral range.
5 The spectral interval carrying the lower cross-validation error compared to the full spectra model are retained.
6 The final PLS2R model is recalibrated with the retained intervals using Venetian-blind cross-validation (10 random blocks).

In this study, to find the optimal interval size, intervals in the range of 5-15 in step of 1 were explored. The iPLS2R was implemented in MATLAB 2017b, Natick, USA, using the freely available iPLS2R toolbox (http://www.models.life.ku.dk/iToolbox).

#### 2.4.4. CovSel

The covariate selection (CovSel) is a popular chemometric technique for filtering important variables (Roger et al., 2011). The CovSel has the advantage over other chemometric techniques in that it can perform simultaneous variable selection for multi-responses. Furthermore, the CovSel has the benefit to be fast and easy to optimize. The background idea of the CovSel is like PLS2R in involving the selection of wavelengths based on response variables followed by orthogonalization steps to remove the variability already explained by the selected wavelengths. In the CovSel, variables are selected one at a time resulting in plots explaining the variance being captured in the function of the number of variables selected. The wavelengths selected from the CovSel are later used for MLR modelling. CovSel was used for simultaneous wavelengths selection for MC and SSC prediction.

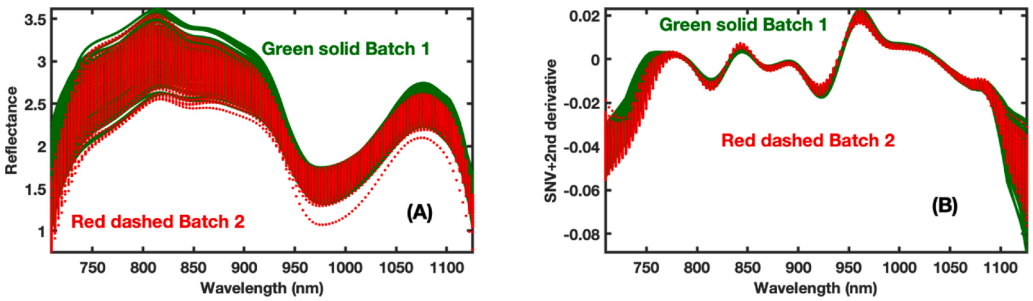


**Fig. 1.** Near-infrared (NIR) reflectance spectra (700-1135 nm) of fruit from two batches (Batch 1 in green solid lines and Batch 2 in red dashed lines). (A) raw reflectance data, and (B) standard normal variate (SNV) followed by 2nd derivative pre-processed spectra.

#### 2.4.5. Model updating with recalibration using new sample

To improve the model performance on the different batch, model updating with recalibration was performed. The model updating was achieved by incorporating a few samples from the new batch into the old batch and recalibration. The model updating was explored for three different sample numbers i.e. incorporating 20 samples, 10 samples and 5 samples. This was done to have an idea about the minimum number of samples that should be sufficient to improve the model performance. The samples to incorporate were selected utilising the Kennard-Stone (KS) samples partition technique (Kennard and Stone, 1969). The model updating was explored in combination with the variable selection both with the iPLS2R and the CovSel. The results of the model updating are provided as the $Q^2$, RMSEP and prediction bias. All data analyses were performed in MATLAB 2017b (Natwick, MA, USA).

## 3. Results

### 3.1. Spectra of pear fruit

Fig. 1 shows the raw reflectance and pre-processed spectra of pear fruit from two batches i.e., Batch 1 (Red dashed lines) and Batch 2 (Blue solid lines). The spectra are presented in the spectral range of 709-1125 nm as the region is dominated by the 3rd overtones of C-H and O-H bonds as useful for prediction of MC and SSC in fresh fruit. The raw reflectance spectra have mainly two dominated peaks around 820 nm and 1080 nm (Fig. 1A). The 2nd derivative revealed further underlying peaks (Fig. 1B).

### 3.2. Reference analysis

A summary of reference measurements for Batch 1 and 2 is presented in Table 1. The reference MC values for Batch 1 and Batch 2 were in the range of 81.09- 88.25% and 80.12 – 87.23 % respectively. The reference SSC for Batch 1 and Batch 2 were in the range of 8.40 – 16.10 % and 9.70–16.40 % respectively. The means for both the batches were similar.

### 3.3. Partial least-squares regression on full spectra (700-1135 nm)

Fig. 2 shows the PLS2R modelling performed without variable selection and model updating. Fig. 2A shows the RMSEC and RMSECV in the function of LVs simultaneously extracted for SSC and MC. A total of 6 LVs were selected for simultaneous prediction of SSC and MC, this was based on noting no further significant decrease in the error with addition of new LVs. The calibration plots for MC and SSC are shown in Fig. 2B and Fig. 2C. $Q^2$'s of 0.81 and 0.70 were noted for MC and SSC, respectively. In comparison to RMSEC, the RMSEP increased for the MC and SSC predictions, respectively. Further, there were high prediction biases for both MC and SSC predictions, indicating that the model trained on Batch 1 is sub-optimal to be applied to Batch 2. This high bias is also visible as the two parallel clouds corresponding to Batch 1 and Batch 2 in Fig. 3B-C. The bias was higher in the case of MC prediction compared to SSC.

### 3.4. Interval selection with iPLS2R

A summary of intervals explored for iPLS2R and the corresponding results are presented in Table 2. In the case of MC prediction, an interval size of 10 was identified as best, leading to an increase in $Q^2$ (from 0.81 to 0.84), a reduction in RMSEP (from 1.44 % to 0.58 %) and reduction in prediction bias (from 1.31 % to 0.19 %) compared to the PLS2R performed without interval selection. In the case of SSC prediction, an interval size of 8 was identified as best leading to an increase in $Q^2$ (from 0.70 to 0.71), reduction in RMSEP (from 0.90 % to 0.63 %) and reduction in prediction bias (from -0.62 % to 0.07 %) compared to the PLS2R performed without interval selection. The best model was selected based on high $Q^2$, and low RMSEP and prediction bias.

The best models with the selected intervals are shown in Fig. 3. In the case of MC, two spectral regions i.e., 743-779 nm and 879-939 nm, were selected corresponding to the model with best performance (Fig. 3A). This model with the selected regions required 4 LVs (Fig. 3B) to lead to the final model (Fig. 3C). In the case of SSC, two spectral regions i.e., 709-759 nm and 789-999 nm, were selected corresponding to the model with the best performance (Fig. 3D). This model with the selected regions required 5 LVs (Fig. 3E) to lead to the final model (Fig. 3F). The green and the red points in the Fig. 3C and 3 F are the samples from Batch 1 and 2 respectively. In summary, the spectral region selection prior to PLS2R improved the model performance i.e., high $Q^2$, low RMSEP and bias.

### 3.5. Discrete wavelength selection with CovSel

Fig. 4A shows the explained variance in NIR spectroscopy data and the simultaneously explained variance in the MC and SSC. In total, 8 wavelengths were selected based on the stability of variance explained in Fig. 4A. The selected bands for simultaneous prediction of MC and SSC were 736, 709, 961, 1109, 1125, 816, 912, 879 nm. The wavelengths are arranged based on decreasing co-variance. The calibration models with selected wavelengths for the MC and SSC are shown in Fig. 4C and D respectively. The different between two lines is the model bias. In the case of the MC, the $Q^2$ was increased (from 0.81 to 0.84), RMSEP was reduced (from 1.44 % to 0.64 %) and prediction bias was reduced (from 1.31 % to -0.29 %) compared to the PLS2R performed without CovSel wavelength selection. In the case of the SSC, the $Q^2$ was increased (from 0.70 to 0.73) but the RMSEP and prediction bias were not improved compared to the PLS2R performed without CovSel wavelength selection.

### 3.6. Model updating with recalibration and iPLS2R modelling

Model updating in general improved the performance of the both PLS2R and iPLS2R modelling in terms of higher $Q^2$ and lower RMSEP and prediction bias, however, the performance of iPLS2R was better for both MC and SSC. In the case of model updating with 20 new samples, intervals of 9 and 11 were identified by iPLS2R as the best for predicting MC and SSC, respectively (Table 3). To MC prediction, the iPLS2R in comparison to PLS2R after model updating increased the $Q^2$ (from 0.85 to 0.87) and reduced the RMSEP (from 0.55 % to 0.52 %) and prediction bias (from 0.16 % to 0.10 %). To SSC prediction, the iPLS2R in comparison to PLS2R after model updating increased the $Q^2$ (0.74 to 0.75) and reduced the RMSEP (from 0.60 % to 0.58 %) and prediction bias (0.08 % to 0.06 %).

In the case of the model updating with 10 new samples, an interval size of 9 and 10 were identified as the best for predicting MC and SSC respectively (Table 3). To MC prediction, the iPLS2R in comparison to PLS2R after model updating increased the $Q^2$ (from 0.83 to 0.86) and

**Table 1**
A summary of destructive measurements performed for moisture and soluble solids content.

| Batch | Moisture content (%) | | | | Soluble solids content (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Std | Min | Max | Mean | Std |
| Batch 1 | 81.09 | 88.25 | 84.61 | 1.37 | 8.40 | 16.10 | 12.80 | 1.31 |
| Batch2 | 80.12 | 87.23 | 84.26 | 1.38 | 9.70 | 16.40 | 12.74 | 1.17 |

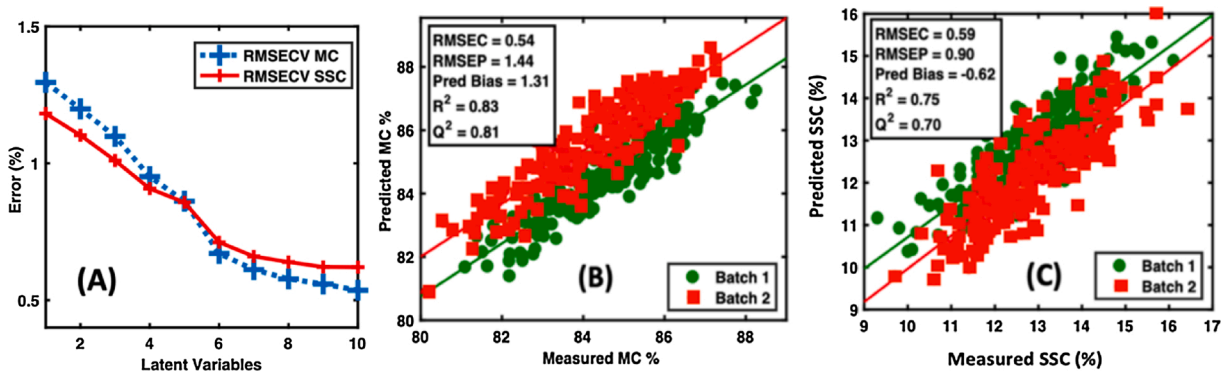


**Fig. 2.** Partial least-squares regression (PLS2R) modelling on complete spectra. (A) Error plot for latent variables (LVs) selection for moisture content (dashed blue) and soluble solids content (Solid red) (6 LVs used selected), (B) model calibration (green circles) and test (red squares) for moisture content and (C) soluble solids content (%).

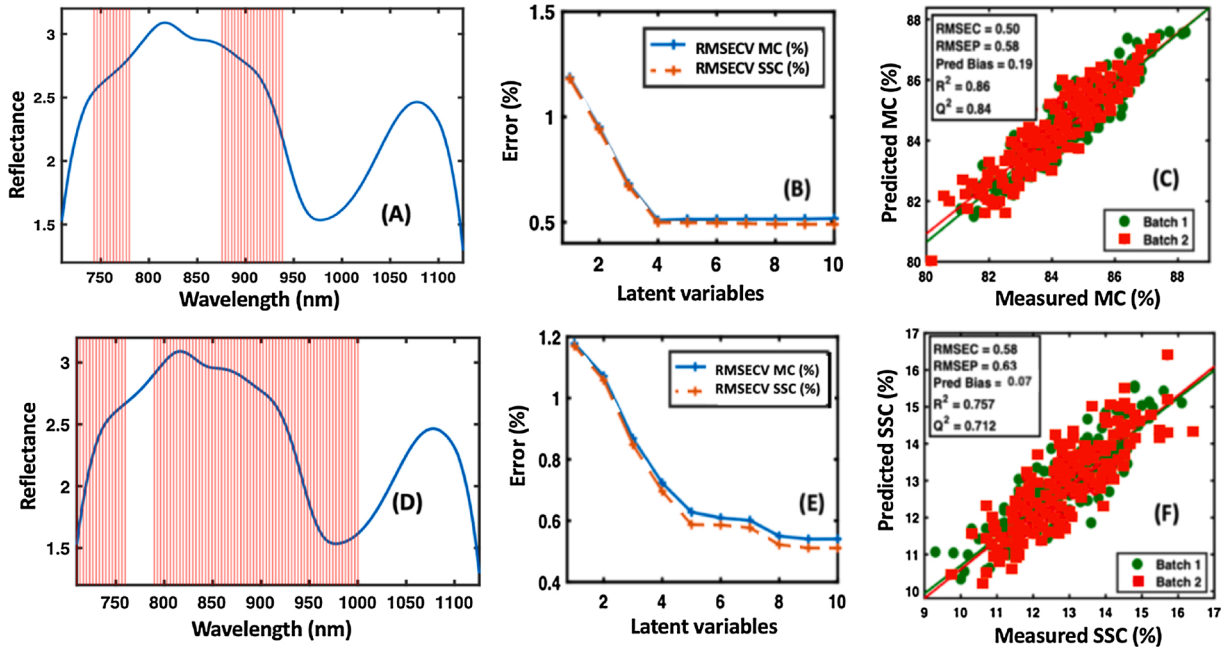


**Fig. 3.** Summary of interval partial least-squares regression (iPLS2R) models for moisture content (MC) and soluble solids content (SSC) prediction. Selected regions for MC (%) (A) and SSC (%) (D), latent variables optimization for MC (B) and SSC (E), model calibration (green circles) and test (red squares) for MC (C) and SSC (F).

**Table 2**
Summary of model for moisture content (MC) and soluble solids content (SSC) prediction with varying intervals of interval partial least-squares regression (iPLS2R). An interval size of 10 was identified for the MC (%) and of size 8 was identified for the SSC (%) prediction. $Q^2$ stands for coefficient of determination for test set and RMSEP stands for root mean squared error of prediction.

| Interval size | Moisture content | | | Soluble solids content | | |
|---|---|---|---|---|---|---|
| | $Q^2$ | RMSEP (%) | Bias (%) | $Q^2$ | RMSEP (%) | Bias (%) |
| 5 | 0.84 | 1.08 | 0.93 | 0.71 | 0.74 | −0.40 |
| 6 | 0.83 | 1.06 | 0.91 | 0.70 | 0.75 | −0.38 |
| 7 | 0.82 | 1.04 | 0.86 | 0.73 | 0.67 | −0.29 |
| 8 | 0.83 | 0.78 | 0.56 | 0.71 | 0.63 | 0.07 |
| 9 | 0.84 | 1.26 | 1.14 | 0.73 | 0.78 | −0.50 |
| 10 | 0.84 | 0.58 | 0.19 | 0.72 | 0.70 | 0.33 |
| 11 | 0.80 | 0.98 | 0.78 | 0.68 | 0.70 | −0.21 |
| 12 | 0.81 | 1.10 | 0.92 | 0.69 | 0.75 | −0.37 |
| 13 | 0.84 | 0.64 | 0.33 | 0.70 | 0.67 | 0.19 |
| 14 | 0.78 | 0.87 | 0.59 | 0.65 | 0.70 | −0.09 |
| 15 | 0.81 | 1.04 | 0.84 | 0.68 | 0.70 | −0.24 |
| Full wavelength PLS2R | | | | | | |
| All bands | 0.81 | 1.44 | 1.31 | 0.70 | 0.90 | −0.62 |

reduced the RMSEP (from 0.64 % to 0.53 %)and prediction bias (from 0.30 % to 0.14 %). To SSC prediction, the iPLS2R in comparison to PLS2R after model updating increased the $Q^2$ (0.75 to 0.76) and reduced the RMSEP (from 0.59 % to 0.57 %) and prediction bias (0.07 % to 0.05 %).

In the case of the model updating with 5 new samples, an interval size of 6 and 15 were identified as the best for predicting MC and SSC respectively (Table 3). In the case of MC prediction, the iPLS2R in comparison to PLS2R after model updating there was no improvement in $Q^2$ and RMSEP, but the prediction bias was reduced from 1.31 % to 0.09 %. In the case of SSC prediction, the iPLS2R in comparison to PLS2R after model updating increased the $Q^2$ (from 0.75 to 0.76) and reduced the RMSEP (from 0.59 % to 0.57 %).

### 3.7. Model updating with recalibration and CovSel variable selection

Like iPLS2R modelling, the CovSel modelling also showed improved prediction with the updated model (Table 4). The best performance was obtained with model updated using 20, followed by 10 and then 5 new samples. Compared to the CovSel modelling for MC prediction without

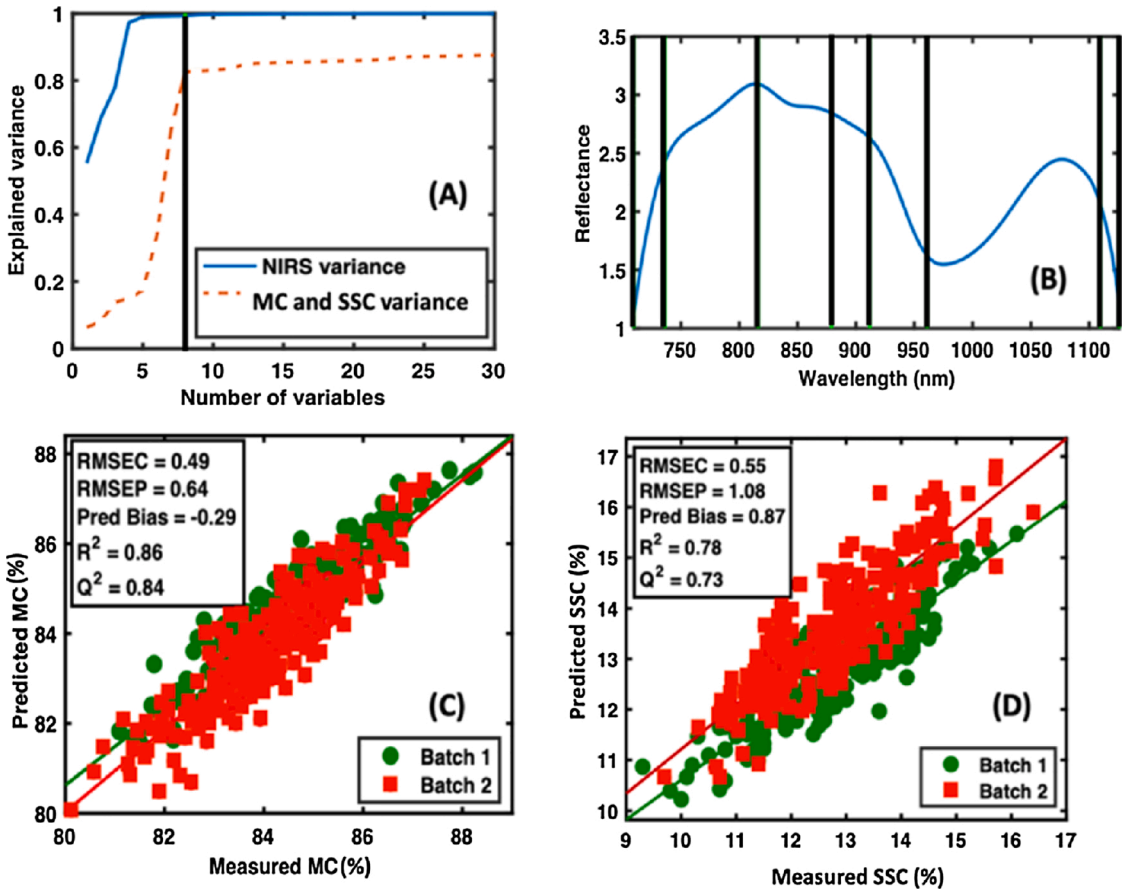


**Fig. 4.** Covariate selection (CovSel) modelling. (A) Variance plot for selecting variables, 8 variables were selected, (B) selected bands (vertical lines), (C) model calibration (green circles) and test (red squares) for moisture content (%) and (D) soluble solids content (%).

**Table 3**
A summary of recalibrated partial least-squares regression (PLS2R) and interval partial least-squares regression (iPLS2R) models with addition of 20, 10 and 5 new samples from new batch. $Q^2$ stands for coefficient of determination for test set and RMSEP stands for root mean squared error of prediction.

| Models (PLS2R and iPLS2R) | Moisture content | | | Soluble solids content | | |
|---|---|---|---|---|---|---|
| | $Q^2$ | RMSEP (%) | Bias (%) | $Q^2$ | RMSEP (%) | Bias (%) |
| Recalibration with 20 new samples | | | | | | |
| PLS2R | 0.85 | 0.55 | 0.16 | 0.74 | 0.60 | 0.08 |
| iPLS2R | 0.87 | 0.52 | 0.10 | 0.75 | 0.58 | 0.06 |
| Recalibration with 10 new samples | | | | | | |
| PLS2R | 0.83 | 0.64 | 0.30 | 0.75 | 0.59 | 0.07 |
| iPLS2R | 0.86 | 0.53 | 0.14 | 0.76 | 0.57 | 0.05 |
| Recalibration with 5 new samples | | | | | | |
| PLS2R | 0.84 | 0.52 | 1.31 | 0.75 | 0.59 | 0.02 |
| iPLS2R | 0.82 | 0.59 | 0.09 | 0.76 | 0.57 | 0.05 |

**Table 4**
A summary of multi-linear regression (MLR)corrected models made on CovSel selected variables for prediction MC and SSC by incorporating new samples in the calibration set. $Q^2$ stands for coefficient of determination for test set and RMSEP stands for root mean squared error of prediction.

| CovSel | Moisture content | | | Soluble solids content | | |
|---|---|---|---|---|---|---|
| | $Q^2$ | RMSEP (%) | Bias (%) | $Q^2$ | RMSEP (%) | Bias (%) |
| No new sample | 0.84 | 0.64 | −0.29 | 0.73 | 1.08 | 0.87 |
| 20 samples from batch 2 | 0.85 | 0.56 | 0.15 | 0.77 | 0.57 | 0.04 |
| 10 samples from batch 2 | 0.84 | 0.57 | 0.14 | 0.77 | 0.58 | 0.09 |
| 5 samples from batch 2 | 0.85 | 0.57 | 0.20 | 0.77 | 0.62 | 0.25 |

model update, the $Q^2$ was increased (from 0.84 to 0.85) and the RMSEP (from 0.64 % to 0.56 %) and the prediction bias (from -0.29 % to 0.15 %) were decreased. Compared to the CovSel modelling for SSC prediction without model updating, the $Q^2$ was increased (from 0.73 to 0.77) and the RMSEP (from 1.08 % to 0.57 %) and the prediction bias (from 0.87 % to 0.04 %) were decreased. In summary, the CovSel improved the model performance and combined with model updating with data from new samples it further reduced the RMSEP and prediction bias.

## 4. Discussion

NIR spectroscopy models of fresh fruit lack robustness when tested in a new batch of fruit measured in a different physical, chemical and environmental conditions (Roger et al., 2003). This problem has been highlighted in multiple scientific pieces of literature (Mishra et al., 2020; Nicolai et al., 2007; Saeys et al., 2019; Walsh et al., 2020), however, a clear solution to the problem is still lacking. In the present work, the use of variable selection has been demonstrated for building robust NIR spectroscopy models that work well on a different batch. Further, model updating with new samples and its combination with the variable selection is demonstrated to gain further improvement in the robustness. The study showed that the models developed on selected regions/specific wavelengths can drastically improve the model performance and it works well when used on a new batch. Further, in combination with model updating with a few new samples, a further reduction in bias and RMSEP was obtained.

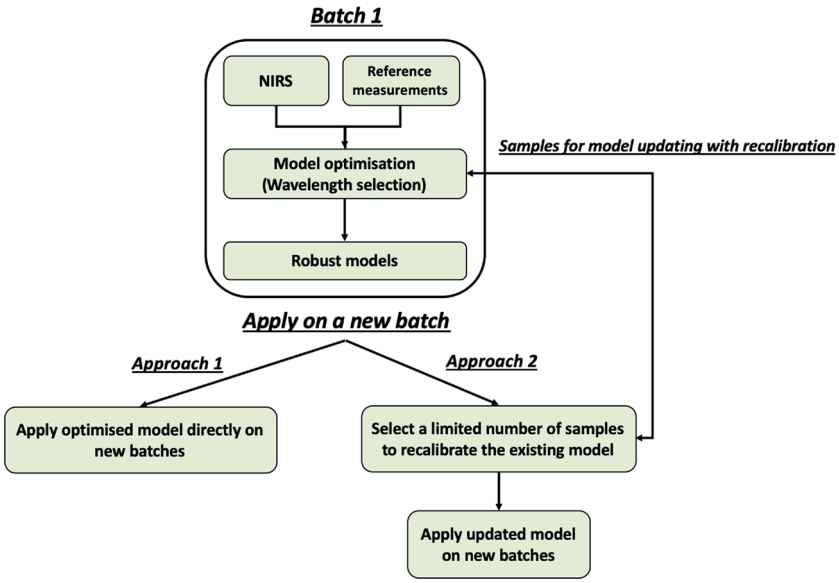


**Fig. 5.** Proposed near-infrared spectroscopy (NIRS) data modelling strategy for robust models and application to a new batch.

In literature, several studies were performed related to the prediction of MC and SSC and the problem of model robustness is very persistent in the case of pear fruit. In a study related to SSC predicting in pear fruit with the absorption and scattering affects a good calibration $R^2$ was obtained but when tested on a validation set the performance was decreased by 50 % (He et al., 2016). In the present work, a similar decrease in performance was observed with the standard PLS2R modelling, but with variable selection, the model performance for the new batch was similar as of the calibration set. This indicates that the models learned with key spectral regions or wavelengths are optimal. Spectral region selection or key wavelengths identification is a logical step as for a fruit property-specific NIR spectral region capture the overtones. For example: to predict MC it is optimal to use the spectral region or key wavelengths that capture the OH bond overtones, rather using the complete spectral range. There are also several other studies on pear fruit quality prediction, but they mostly lack either a new batch (Sun et al., 2009; Wang et al., 2017; Yu et al., 2018; Yuan et al., 2020) or the modelling only involved calibration and cross-validation step and no independent test step (Adebayo et al., 2017; Travers et al., 2014). To have a better understanding of model robustness it is highly recommended to always perform multi-batch experiments in NIR spectroscopy related to fruit.

In this work, without model updating with new samples, the interval-based approach (iPLS2R) worked better compared to the filter-based method (CovSel) in terms of high $Q^2$, and low RMSEP and prediction bias. However, when the model was updated with a few new samples than both the variable selection approaches showed similar performance in-term of $Q^2$ and the RMSEP. The prediction bias was lower for the model made on selected spectral regions with iPLS2R compared to individual wavelengths selected with the CovSel. The better performance of interval-based approach can be linked to the basics of NIR spectroscopy as the NIR spectroscopy data is made of underlying peaks which are expressed over specific spectral regions rather than a single wavelength. Based on the physical, chemical and environmental conditions these peaks might shift or increase or decrease in intensity (Roger et al., 2003; Zeaiter et al., 2006). In such cases, a model based on the selected single band might lead to over or underestimation of property to be predicted, whereas a model with the selected region should be more robust. A point of key importance is to explore the interval size in the case of iPLS2R as different underlying peaks can be of different width and using a single interval size might lead to sub-optimal spectral region selection.

Model updating with a few new samples has the main benefit of reducing the prediction bias. However, it is always challenging to decide on how many samples are required from the new batch. In this work, three different samples sizes (5, 10 and 20) were explored in combination with variable selection. The results showed that 5 samples were not enough to improve the model performance as the RMSEP and prediction bias were higher, but 10 samples were sufficient to keep the RMSEP and prediction bias low. Using 20 samples could improve the performance further, but it is generally better to have minimum new samples to reduce the time and work on new measurements and reap the complete benefit of NIR spectroscopy. On other hands, if a slightly higher RMSEP and prediction bias are allowed then the iPLS2R modelling without model updating with new samples is sufficient to predict MC and SSC in a new batch. Based on the findings from this study, we suggest a NIR spectroscopy modelling procedure for robust NIR spectroscopy data modelling and for using it on new batches. The procedure is shown in Fig. 5. The methodology suggests that for a single batch the model optimisation should be performed with variable selection. Later, to use it on new batch either the optimised model can be directly applied on new samples, or a few selected samples can be used for model updating and later the updated model can be used on the new batch.

## 5. Conclusion

Pear fruit quality parameter such as MC and SSC are key parameters used to decide on harvest date or effect of postharvest storage conditions. NIR spectroscopy is widely explored for that purpose but NIR spectroscopy models often fails when used in a new batch such as samples from a new season, a new cultivar and if samples are measured under different temperature conditions. The results from this study showed that a combination of variable selection and model updating allowed development of robust NIR spectroscopy models for pear fruit that works well when used in a different batch. The results showed that developing models with the key spectral ranges and wavelengths are more robust compared to the standard PLS2R modelling using complete spectra. The CovSel showed less improvement compared to iPLS2R approach. In the case of MC prediction, the variable selection (iPLS2R) reduced the bias (from 1.31 % to 0.19 %) and the RMSEP (from 1.44 % to 0.58 % in) compared to the standard PLS2R. In the case of SSC prediction, the variable selection (iPLS2R) reduced the bias (from -0.62 % to 0.07 %) and the RMSEP (from 0.90 % to 0.63 %) compared to the standard partial least-squares regression. Further, the model updating with recalibration using just 10 new samples drastically reduced the

RMSEP and prediction bias, for both MC and SSC prediction. In the case of MC, spectral regions of 749-759 nm and 879-939 nm were identified as the most important region. In the case of the SSC, 709-759 nm and 789-999 nm were found to be important spectral regions. To develop robust models for NIR spectroscopy related to fruit, it is highly recommended that PLS2R models should be optimised with the use of variable selection methods. If variable selection does not improve model performance on new batch, then model should be updated with a few samples from the new batch. The spectral regions identified in this work can also be used in other studies which uses similar instrument. The presented approach should not be limited to pear fruit but in general the spectral region selection and key wavelengths identification can support robust NIR spectroscopy model development.

## CRediT authorship contribution statement

**Puneet Mishra:** Conceptualization, Data curation, Investigation. **Ernst Woltering:** Writing - review & editing. **Bastiaan Brouwer:** Formal analysis, Software, Visualization. **Esther Hogeveen-van Echtelt:** Writing - review & editing.

## Declaration of Competing Interest

None.

## Acknowledgments

Financial support was granted through Foundation TKI, The Netherlands Horticulture and Starting Materials and other private partners in the project (TU-16025 (1605-043) Humistatus). Additionally, we would like to thank Mariska Nijenhuis, Manon Mensink, Najim El Harchioui, Frank van de Geijn, and Marcel Staal for their help in sample preparation and analysis for MC and SSC.

## References

Adebayo, S.E., Hashim, N., Hass, R., Reich, O., Regen, C., Munzberg, M., et al., 2017. Using absorption and reduced scattering coefficients for non-destructive analyses of fruit flesh firmness and soluble solids content in pear (Pyrus communis' Conference')-An update when using diffusion theory. Postharvest Biol. Technol. 130, 56–63. https://doi.org/10.1016/j.postharvbio.2017.04.004.

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. https://doi.org/10.1366/0003702894202201.

Bobelyn, E., Serban, A.-S., Nicu, M., Lammertyn, J., Nicolai, B.M., Saeys, W., 2010. Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. Postharvest Biol. Technol. 55 (3), 133–143. https://doi.org/10.1016/j.postharvbio.2009.09.006.

He, X.M., Fu, X.P., Rao, X.Q., Fang, Z.H., 2016. Assessing firmness and SSC of pears based on absorption and scattering properties using an automatic integrating sphere system from 400 to 1150 nm. Postharvest Biol. Technol. 121, 62–70. https://doi.org/10.1016/j.postharvbio.2016.07.013.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. Technometrics 11 (1), 137–148. https://doi.org/10.1080/00401706.1969.10490666.

Lin, H., Ying, Y., 2009. Theory and application of near infrared spectroscopy in assessment of fruit quality: a review. Sens. Instrum. Food Qual. Saf. 3 (2), 130–141. https://doi.org/10.1007/s11694-009-9079-z.

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 https://doi.org/10.1016/j.postharvbio.2019.111003.

Mehmood, T., Liland, K.H., Snipen, L., Sæbø, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemometr. Intell. Lab. Syst. 118, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.

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

Mishra, P., Asaari, M.S.M., Herrero-Langreo, A., Lohumi, S., Diezma, B., Scheunders, P., 2017. Close range hyperspectral imaging of plants: A review. Biosyst. Eng. 164, 49–67. https://doi.org/10.1016/j.biosystemseng.2017.09.009.

Mishra, P., Roger, J.M., Rutledge, D.N., Woltering, E., 2020. SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. Postharvest Biol. Technol. 168, 111271 https://doi.org/10.1016/j.postharvbio.2020.111271.

Nascimento, P.A.M., Carvalho, L. C. d., Júnior, L.C.C., Pereira, F.M.V., Teixeira, G. H. d. A., 2016. Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR). Postharvest Biol. Technol. 111, 345–351. https://doi.org/10.1016/j.postharvbio.2015.08.006.

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. https://doi.org/10.1016/j.postharvbio.2007.06.024.

Nordey, T., Joas, J., Davrieux, F., Chillet, M., Lechaudel, M., 2017. Robust NIRS models for non-destructive prediction of mango internal quality. Sci. Hortic. 216, 51–57. https://doi.org/10.1016/j.scienta.2016.12.023.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval Partial Least-Squares Regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54 (3), 413–419. https://doi.org/10.1366/0003702001949500.

Palmer, J.W., Harker, F.R., Tustin, D.S., Johnston, J., 2010. Fruit dry matter concentration: a new quality metric for apples. J. Sci. Food Agric. 90 (15), 2586–2594. https://doi.org/10.1002/jsfa.4125.

Peirs, A., Tirry, J., Verlinden, B., Darius, P., Nicolaï, B.M., 2003. Effect of biological variability on the robustness of NIR models for soluble solids content of apples. Postharvest Biol. Technol. 28 (2), 269–280. https://doi.org/10.1016/S0925-5214(02)00196-5.

Rady, A.M., Guyer, D.E., 2015. Evaluation of sugar content in potatoes using NIR reflectance and wavelength selection techniques. Postharvest Biol. Technol. 103, 17–26. https://doi.org/10.1016/j.postharvbio.2015.02.012.

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. Chemometr. Intell. Lab. Syst. 66 (2), 191–204. https://doi.org/10.1016/S0169-7439(03)00051-0.

Roger, J.M., Palagos, B., Bertrand, D., Fernandez-Ahumada, E., 2011. CovSel: Variable selection for highly multivariate and multi-response calibration Application to IR spectroscopy. Chemometr. Intell. Lab. Syst. 106 (2), 216–223. https://doi.org/10.1016/j.chemolab.2010.10.003.

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 doi:UNSP 11098110.1016/j.postharvbio.2019.110981.

Sun, T., Lin, H.J., Xu, H.R., Ying, Y.B., 2009. Effect of fruit moving speed on predicting soluble solids content of' Cuiguan' pears (Pomaceae pyrifolia Nakai cv. Cuiguan) using PLS and LS-SVM regression. Postharvest Biol. Technol. 51 (1), 86–90. https://doi.org/10.1016/j.postharvbio.2008.06.003.

Teh, S.L., Coggins, J.L., Kostick, S.A., Evans, K.M., 2020. Location, year, and tree age impact NIR-based postharvest prediction of dry matter concentration for 58 apple accessions. Postharvest Biol. Technol. 166, 111125 https://doi.org/10.1016/j.postharvbio.2020.111125.

Travers, S., Bertelsen, M.G., Petersen, K.K., Kucheryavskiy, S.V., 2014. Predicting pear (cv. Clara Frijs) dry matter and soluble solids content with near infrared spectroscopy. Lwt-Food Sci. Technol. 59 (2), 1107–1113. https://doi.org/10.1016/j.lwt.2014.04.048.

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 https://doi.org/10.1016/j.postharvbio.2020.111139.

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. https://doi.org/10.3390/s150511889.

Wang, J.H., Wang, J., Chen, Z., Han, D.H., 2017. Development of multi-cultivar models for predicting the soluble solid content and firmness of European pear (Pyrus communis L.) using portable vis-NIR spectroscopy. Postharvest Biol. Technol. 129, 143–151. https://doi.org/10.1016/j.postharvbio.2017.03.012.

Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58 (2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Yu, X.J., Lu, H.D., Wu, D., 2018. Development of deep learning method for predicting firmness and soluble solid content of postharvest Korla fragrant pear using Vis/NIR hyperspectral reflectance imaging. Postharvest Biol. Technol. 141, 39–49. https://doi.org/10.1016/j.postharvbio.2018.02.013.

Yuan, L.M., Mao, F., Chen, X.J., Li, L.M., Huang, G.Z., 2020. Non-invasive measurements of' Yunhe' pears by vis-NIRS technology coupled with deviation fusion modeling approach. Postharvest Biol. Technol. 160 doi:UNSP11106710.1016/j.postharvbio.2019.111067.

Zeaiter, M., Roger, J.M., Bellon-Maurel, V., 2006. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. Chemometr. Intell. Lab. Syst. 80 (2), 227–235. https://doi.org/10.1016/j.chemolab.2005.06.011.

Zou, X.B., Zhao, J.W., Povey, M.J.W., Holmes, M., Mao, H.P., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667 (1–2), 14–32. https://doi.org/10.1016/j.aca.2010.03.048.