FRUITNIR-GUI: A graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction

Puneet Mishra, Jean-Michel Roger, Federico Marini, Alessandra Biancolillo, Douglas N Rutledge

▶ **To cite this version:**

Puneet Mishra, Jean-Michel Roger, Federico Marini, Alessandra Biancolillo, Douglas N Rutledge. FRUITNIR-GUI: A graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction. Postharvest Biology and Technology, 2021, 175, pp.111414. 10.1016/j.postharvbio.2020.111414 . hal-03040830

HAL Id: hal-03040830

https://agroparistech.hal.science/hal-03040830

Submitted on 31 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Postharvest Biology and Technology 175 (2021) 111414

Contents lists available at ScienceDirect

Postharvest Biology and Technology

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

FRUITNIR-GUI: A graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction

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

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

ARTICLE INFO

Keywords:
Chemometrics
User-interface
Non-destructive
Fruit quality

ABSTRACT

Near infrared (NIR) spectroscopy is widely used for non-destructive prediction of fruit traits. Common traits such as dry matter (DM) and soluble solids contents (SSC) can be predicted with reliable accuracy. However, the main problem with NIR spectroscopy is that a model developed on one batch may not perform very well when tested on other batches. Reasons for that are the physical, chemical and environmental differences between the experiments performed in different batches. To deal with these issues, approaches such as variables selection, dynamic orthogonal projection (DOP) and transfer component analysis (TCA) can be used. However, the techniques are known but it is rarely possible for a new user or non-specialist to implement them in the practical situations. To overcome this limitation, for the first time, a graphical user interface-based toolbox (FRUITNIR-GUI) for basic chemometric data processing (regression and variable selection) is developed and presented. The GUI allows performing model adaption and maintenance in the context of multi-batch NIR spectroscopic experiments related to fruit. Furthermore, a case-study demonstrating its effectiveness in correcting for seasonality when predicting DM in apples is presented. The toolbox provides a push-button approach to build chemometric models of varying complexity for the characterization of fruit quality. Moreover, approaches such as variable selection and batch correction with DOP and TCA can improve the model performances on new batches. FRUITNIR-GUI can be freely downloaded at https://github.com/puneetmishra2/FRUITNIR and run using the password "welovenirs" (without quotation marks).

1. Introduction

Near-infrared (NIR) spectroscopy is the most popular non-destructive sensing approach for rapid assessment of fruit properties (Nicolai et al., 2007; Sun et al., 2020). Indeed, NIR spectroscopy is based on the vibrational combinations and overtones of several fundamental bonds such as OH, NH and CH which can be correlated to fruit properties (Mishra et al., 2017, 2020a). Among the various properties, DM and SSC are widely explored for deciding on optimum harvest dates and in fruit sorting lines (Mishra et al., 2021; Sun et al., 2020; Walsh et al., 2020a).

NIR spectroscopy is widely used for the estimation of fruit properties; however, a major challenge with the technique is related to the inaccuracy of models when a relevant variability related to batch effects is present in the data (Nicolai et al., 2007; Sun et al., 2020). Indeed, NIR users often complain that the models do not perform well when used on data collected on fruit harvested in a different season, that a new instrument may require a new calibration or that temperature affects the predictions. In the domain of NIR spectroscopy, these problems are well known and can arise because of a wide range of physical, chemical and environmental factors (Zeaiter et al., 2006). Differences in the spectroscopic signatures can also derive from instrumental characteristics, such as temperature variations due to long use of the infrared light source or degradation in sensor detectivity. Approaches such as variable selection, batch effect correction with techniques such as dynamic orthogonal

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

https://doi.org/10.1016/j.postharvbio.2020.111414
Received 23 July 2020; Received in revised form 10 November 2020; Accepted 11 November 2020
Available online 26 November 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/).

projection (DOP) (Zeaiter et al., 2006) and transfer component analysis (TCA) (Pan et al., 2011) can improve the predictive performances of models when used on a new batch (Mishra et al., 2020b).

However, here it should be stressed that, although several approaches to correct for the differences between the batches are available (Mishra et al., 2020b), it is rarely possible for a new user or non-specialist to implement these techniques in practical situations. Indeed, several scientific toolboxes are freely available for the chemometric analysis of multivariate data (Daszykowski et al., 2007; Mishra et al., 2020c; Mobaraki and Amigo, 2018), but none of them are capable to cope with the challenges related to batch effects when characterizing fruit quality. To deal with this issue, the present work provides a graphical user interface-based toolbox (FRUITNIR-GUI) for basic chemometric processing (regression and variable selection), with the possibility of performing model adaption and maintenance, so as to be applicable to multi-batch NIR spectroscopic experiments related to fresh fruit. Furthermore, a case-study is presented demonstrating its effectiveness in correcting for seasonality when predicting DM in apples.

2. Material and methods

2.1. Software description

The FRUITNIR-GUI was built utilising the application builder in MATLAB version 2018b (The Mathworks, Natick, MA, USA). All the functions included in the toolbox are either built-in in the programming environment or codes developed in-house. The application can be downloaded and installed in MATLAB (preferred versions: 2018b or more recent), can be run through the '.mlapp' files in the MATLAB command line or can be used as a stand-alone executable. If users do not have a MATLAB version recent enough (from 2018b onwards), it is recommended to install the free MATLAB runtime tool and run the app as standalone. All the executables and MATLAB functions can be downloaded from (https://github.com/puneetmishra2/FRUITNIR). In the GitHub repository, the standalone toolbox executable files can be downloaded as *'FRUITNIR.zip'*(https://github.com/puneetmishra2/FRUITNIR/raw/master/FRUITNIR.zip) and the function for running the tools in command line as '*Fruitnir_functions.zip*'. The dataset corresponding to the case-study discussed in this article can be obtained from the publisher of original data set (Teh et al., 2020). All the files are available at the link (https://github.com/puneetmishra2/FRUITNIR). To run the toolbox from the command line, users should use the toolbox folder as the current folder and type T1 on the command line, so to start the main graphical user interface. The users should input the password: 'welovenirs' (without quote marks) and click run. Then, users can load data and run the analysis. The GUI supports. csv, .xlsx and. mat data formats. A summary of the software architecture is presented in Fig. 1.

The toolbox has options for loading data, three levels of pre-processing, i.e. smoothing, scatter correction and normalisation, and differentiation. Additionally, the toolbox has options for partial least-squares (PLS) regression, covariate selection (CovSel) for variable selection, model maintenance by dynamic orthogonal projections (DOP) and domain adaption with transfer component analysis (TCA). For performance comparison of models two different statistical parameters are integrated in the GUI i.e. coefficient of determination of prediction (R^2_P) (Eq. 1), and the root mean squared error of prediction error (RMSEP) (Eq. 2).

$$R_P^2 = r(y, \hat{y})^2 \quad (1)$$

$$RMSEP = \sqrt{\sum \frac{(Y - \hat{Y})^2}{n}} \quad (2)$$

Where, r is the correlation coefficient, Y is the expected value, $\hat{Y}$ is the predicted values and n is the total number of samples.

2.2. Dataset for demonstrating the use of FRUITNIR-GUI

The demonstration of the GUI was performed on a desktop computer equipped with a 3.60 GHz Intel® Xeon® W-2133 processor (Intel Corporation, Santa Clara, CA) and 64 GB RAM, running Microsoft Windows 10 operating system (Microsoft Inc., Redmond, WA) at 64-bit and MATLAB 2018b (The Mathworks, Natick, MA). To demonstrate the functionality of the GUI, a dataset related to the prediction of DM in apples harvested in two different seasons was used (Teh et al., 2020). As explained in the original work (Teh et al., 2020), the NIR and DM measurements were carried out in 2015 and 2016 and involved assessment of 2252 fruits from 58 accessions at three orchard sites of Washington State University apple breeding program (WABP). The 58 accessions included 34 WABP apple selections and five commercial cultivars (i.e., Cripps Pink, Fuji, Gala, Golden Delicious and Honeycrisp). As explained in the original work (Teh et al., 2020), fruits were stored at 2 °C for two months. After storage, the samples were stabilized for a week at room temperature (25 °C) and five apples were randomly selected for NIR measurements. NIR spectra were acquired using F-750 Produce Quality Meter (Felix Instruments, Camas, WA, USA). As explained in the original work (Teh et al., 2020), the DM measurements were done by sampling and dehydrating a cylindrical core in a food dehydrator. More specific details on the dataset are summarized in Table 1. In this study, since the aim of the methods implemented in the toolbox is to reduce/eliminate the experimental variability ascribable to the batch effect, the models were built on data from year 2015 (training set) and were tested on year 2016 data. However, since one of the techniques implemented in the toolbox (DOP) requires some data from the new batches to perform the correction, and in order to avoid any

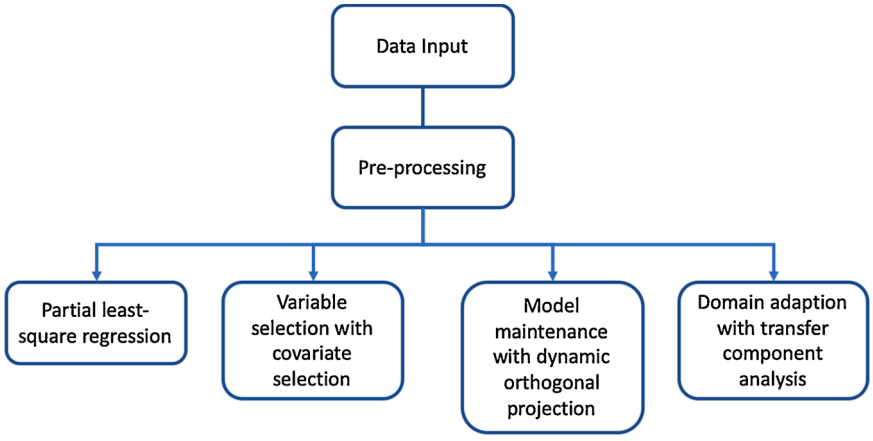

Fig. 1. A summary of the methods available in the toolbox.

Table 1
Description of near-infrared (NIR) and dry matter (DM) for apple dataset. The data for optimizing the dynamic orthogonal projection (DOP) were selected from test data using the Kennard -Stone algorithm.

Dataset	Spectral range (nm)	Training		Optimizing DOP		Testing	
		NIR	DM (mean ± std)	NIR	DM (mean ± std)	NIR	DM (mean ± std)
Apple season correction	729–975	1219 × 83	15.46 ± 1.49	207 × 83	15.66 ± 2.20	800 × 83	15.51 ± 1.84

over optimism due to the use of the same set of samples for model selection/optimization and for validation, the year 2016 data were further split into two subsets: ~20 % of the 2016 data, selected by the Kennard-Stone algorithm, were chosen as the subset to be used to calculate the DOP correction while the remaining ~80 % were left completely out to constitute the external test set.

2.3. Software architecture and brief mathematical background of the techniques available

2.3.1. Pre-processing techniques

Data pre-processing is a major step to clean and homogenise the data prior to data analysis (Mishra et al., 2020d, Mishra et al., 2020e). When pre-processing spectral data, multiple steps, such as smoothing, scatter correction and normalisation or differentiation, may be involved. In the FRUITNIR-GUI, several common pre-processing methods are available.

2.3.1.1. Smoothing operations. In the GUI, several techniques are provided for performing the spectral smoothing. Three window-based smoothing techniques, i.e., Savitzky-Golay (SAVGOL) (Savitzky and Golay, 1964), moving average and moving polynomial are provided. Two data decomposition and reconstruction techniques, i.e., principal components reconstruction and independent components reconstruction are also provided. All the smoothing methods are implemented using the codes presented in (Roger et al., 2020).

2.3.1.2. Scatter correction, baseline correction and normalisation. Multivariate data, and especially spectral data, suffer from a range of physical and chemical effects leading to non-zero baselines, additive and multiplicative effects. These effects also need to be corrected prior to data analysis and this often requires some sort of normalisation. In the toolbox, the users may select several scatter correction and spectral normalisation techniques, including detrending, offset correction, multiplicative scatter correction (Isaksson and Næs, 1988), spline correction, asymmetric least-squares (AsLS) correction (Boelens et al., 2004), standard normal variates (SNV) (Barnes et al., 1989), variable sorting for normalisation (VSN) (Rabatel et al., 2020), probabilistic quotient normalisation (PQN) (Dong et al., 2011), robust normal variates (RNV) (Guo et al., 1999), log transform, autoscaling, 1st derivative, 2nd derivative (Savitzky and Golay, 1964), min-max, norm, range and max correction. All the correction and normalisation methods are implemented using the codes presented in (Roger et al., 2020). Smoothing should be performed before baseline correction and normalisations, as these techniques can be affected by high-frequency noise.

2.3.1.3. Derivatives. Derivatives are used to reveal the underlying peaks (Savitzky and Golay, 1964). FRUITNIR-GUI allows to calculate 1st or 2nd derivatives with pre-defined settings, or to select a specific order of the derivative (together with setting custom values for the meta-parameters, e.g., the degree of the interpolating polynomial and the data point window, if using the Savitzky-Golay algorithm) as a 3rd pre-processing step. The algorithms for this operation are implemented by means of an in-house code ((Roger et al., 2020)).

2.3.2. Partial least-squares regression

PLS regression is a common chemometric technique for calibration problems involving NIR spectroscopic data (Wold et al., 2001). In particular, PLS deals with the multi-collinearity in the multivariate signal by projecting the data onto a subspace of latent variables (LVs) and compressing the relevant information in the **X** block into a few orthogonal scores, which are extracted so to have maximum covariance with the response(s). This guarantees at the same time that the scores are explanatory (i.e., provide a "good summary") of the variance in **X**, and that they are relevant for predicting the response(s) **Y**. A more detailed description of the method can be found in (Geladi and Kowalski, 1986; Wold et al., 2001). In the GUI, PLS models are calculated by means of the MATLAB's 'plsregress', which has been integrated with a function performing a 10-fold cross validation procedure approach is integrated for the selection of the optimal model complexity (number of LVs defining the subspace for data projection).

2.3.3. Covariate selection

Covariance selection (CovSel) is a popular chemometric technique for filtering (selecting) important variables (Roger et al., 2011) in the context of predictive modeling (regression or classification). In CovSel, variable selection is accomplished by iterating two steps: (i) the X-variable having maximum covariance with the response(s) is selected; (ii) both the predictor and the response matrices are orthogonalized with respect to the selected variable. These two steps are repeated until a pre-defined criterion is met: one possibility is to inspect the plot of the explained variation as a function of the number of selected variables and choose the complexity corresponding to a clear inflection point in the graph; another options, which is the most commonly used, is to retain the number of variables which leads to the minimum root mean square error in a cross-validation procedure. In the toolbox, CovSel is implemented by means of an in-house code (Roger et al., 2011). In particular, once the variables are selected as described above, the final calibration model is built using multiple linear regression (MLR), through the built-in MATLAB function 'fitlm'.

2.3.4. Dynamic orthogonal projection

Dynamic orthogonal projection (DOP) is a model maintenance method developed to deal with physical, chemical and environmental affects in spectroscopic modelling (Zeaiter et al., 2006). The approach is based on the correction of the calibration dataset based on the new reference measurements performed in different physical, chemical and environmental conditions. The correction is performed using orthogonal projections based on the subspace defined by the difference of the calibration spectra and the new condition spectra. Let ***r***be a set of samples measured in the new conditions. Let $\mathbf{Y_r}$ be the reference values and $\mathbf{X_r}$ the measured spectra of these samples. The DOP method starts by estimating virtual standards, i.e., the spectra $\mathbf{X_r^*}$ that should have been measured in correspondence with $\mathbf{Y_r}$, if the calibration conditions had not varied. This is accomplished by means of linear combinations of the original calibration data matrix, whose coefficients are calculated using kernels centred on $\mathbf{Y_r}$ values. Once the virtual standards are prepared then the difference spectra between $\mathbf{X_r}$ and $\mathbf{X_r^*}$ are calculated. The orthogonal basis for the difference spectra is estimated by singular value decomposition (SVD) and finally, the original spectra are projected orthogonally to that basis. This removes the external influences from the spectra and then the model recalibrated on these data becomes insensitive to the differences (physical, chemical and environmental conditions). In the GUI, DOP is implemented by means of an in-house function (Mishra et al., 2020b); once the data are corrected, then a PLS regression

model is built as described previously.

2.3.5. Transfer component analysis

TCA applies when data that are expected to have very similar/identical variability, instead show differences in their statistical distributions (Pan et al., 2011). For instance, in the case of NIR spectroscopy, it can happen that, if spectra are collected on two different instruments, or at two different temperatures or, again, during two different seasons, the corresponding variance-covariance matrices will be different as well. TCA aims at finding a latent feature space that minimizes the difference between the distributions resulting from the two data sets, at the same time preserving as much as possible of their respective original variance. In particular, TCA embeds the data from both sources into a shared low dimensional latent space using a nonlinear mapping, defined implicitly by a kernel matrix. In the toolbox, TCA is implemented as explained in (Pan et al., 2011), and once the data from both sources (both batches in the case of NIR applications) are projected onto their common feature space, then standard PLS regression modeling is performed as already described. In the toolbox, TCA is implemented as described in the original publication (Pan et al., 2011) by means of the functions developed by Yan (2020).

2.3.6. Data loading and pre-processing to the GUI

The FRUITNIR-GUI provides the possibility to load two sets of data together. The two datasets can be either a training/test set pair, or they can be data from two seasons, two temperature conditions or two instruments. The GUI for loading the dataset is shown in Fig. 2. The first step is to select the type of file to be loaded. There are currently three file type options, namely. csv, .xlsx and. mat formats; this choice is based on their being among the most popular data output formats in many NIR spectrometers or, in the case of. mat, to allow MATLAB users to directly input their data. Once the data are loaded, the three pre-processing drop-down menus can be used for smoothing, scatter correction and differentiation. It is recommended to follow these steps while loading data: first load the calibration dataset, select the pre-processing strategy and apply it and then load the test dataset; then, the same pre-processing will be applied to the test dataset automatically, once it is loaded. There are four processing options provided in the toolbox, i.e., standard PLS regression, CovSel variable selection, DOP or TCA approaches for the correction of batch effects. When the values of the response variables for the test set are not available, then only TCA can be used for batch effect correction, as it is an unsupervised strategy. On the other hand, when DOP is used to remove the batch effects, the second matrix to be loaded should be the tuning set needed for performing the correction: in such cases, the final test set for model validation can be loaded in a subsequent interface, specifically called by the DOP routine. Once the data are loaded, the user can choose the method from among the four options available through the pop-up menu shown in Fig. 2, and then perform the analysis by pressing the RUN button. In any situation, if the user decides to restart the GUI then the 'Restart' button can be used.

3. Results and discussion

3.1. Partial least-square regression

PLS analysis allows selecting the optimal number of LVs based on cross-validation. The optimization of the model complexity utilizes plots of the explained variance and of the mean squared error (MSE: the mean of the squared residuals, i.e., differences between measured and predicted response) as a function of the number of LVs (Fig. 3). The inflection point in Fig. 3 was used to decide on the number of LVs to be retained in the final model. In the case-study presented, 6 LVs were chosen to build the final model. The results of calibration and test are shown in Fig. 4. PLS regression attained a R^2_P of 0.90 with the error (RMSEP: The root mean squared error of prediction) of 0.80 % (Fig. 4B). The error was higher compared to those attained with the calibration set (Fig. 4A), indicating that a batch effect might be present and could have affected the application of the PLS model when tested on data from a new season. Similar model failure was also highlighted in some recent studies such as in relation to prediction of DM and SSC in pear fruit under different storage conditions, where the standard PLS regression model attained high error (Mishra et al., 2021). Similarly, standard PLS regression models have been reported to lead to high RMSEP in the prediction of DM in mango fruit of different seasons (Mishra and Nikzad-Langerodi, 2020; Mishra et al., 2020b). Previous results and the results obtained in the present study suggest that PLS alone is not

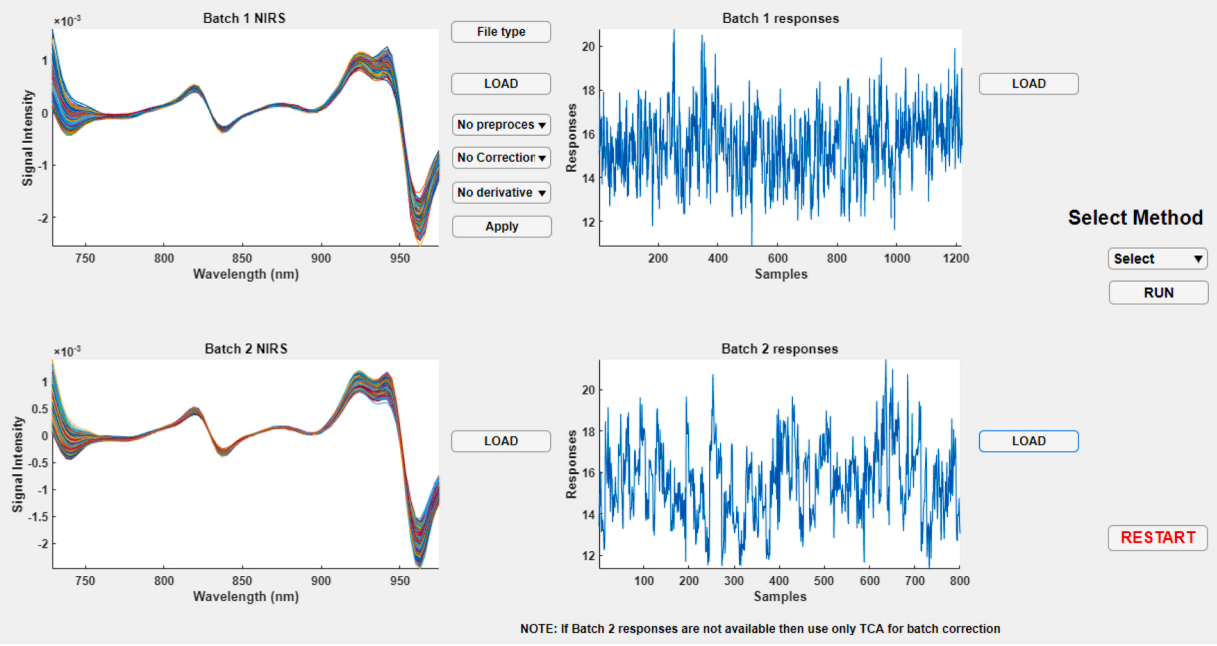

Fig. 2. GUI for loading and pre-processing datasets. Two different batches can be loaded through the same interface. Prior to loading second batch, batch 1 should be pre-processed with the desired pre-processing. The same pre-processing will automatically be applied to the second batch data when its loaded.

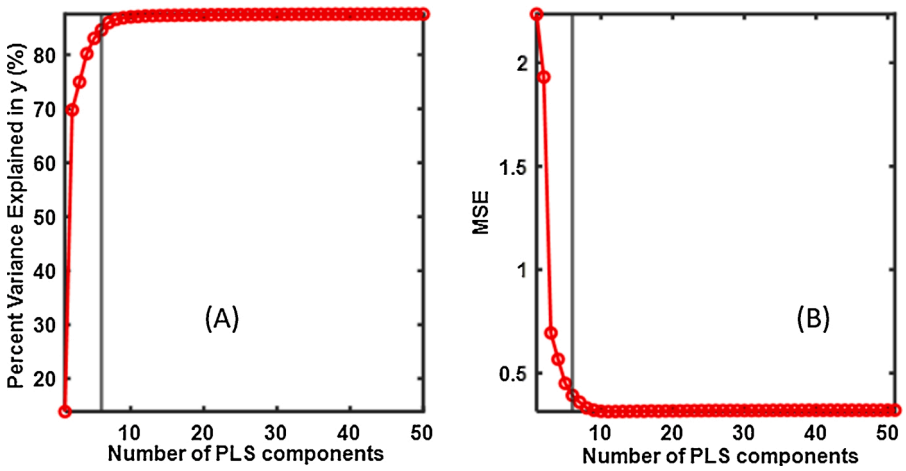

Fig. 3. Error plots for selection of latent variables (LVs). (A). Explained variance in response variables as a function of the number of LVs, and (B). Mean squared error (MSE: the mean of the squared residuals, i.e., differences between measured and predicted response) as function of the number of LVs.

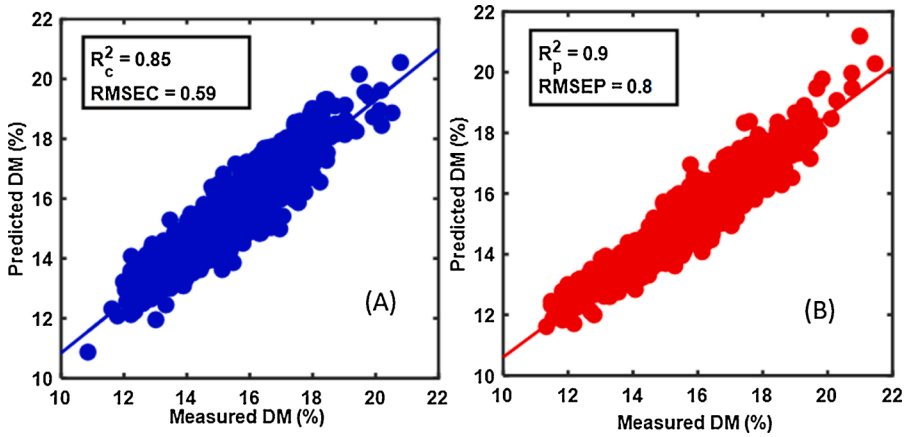

Fig. 4. Partial least-squares (PLS) regression for dry matter (DM %) prediction in apples. Calibration set (A) and test set (B). R^2_c: Coefficient of determination for calibration set, R^2_p: coefficient of determination for test set, RMSEC: root mean squared error of calibration, and RMSEP: root mean squared error of prediction.

capable in dealing with the batch effects. Hence, in the following sections, the benefits resulting from the use of the GUI for correcting the batch effects will be illustrated, and the outcomes will be compared to the performances of standard PLS regression modeling.

3.2. Selecting variables with CovSel

Variable selection allows to generalize the model performance by retaining key variables that relate the most with the property of interest. Fig. 5 shows the results of CovSel analysis performed on the apple dataset. Fig. 5A shows the 20 variables selected by CovSel, most of

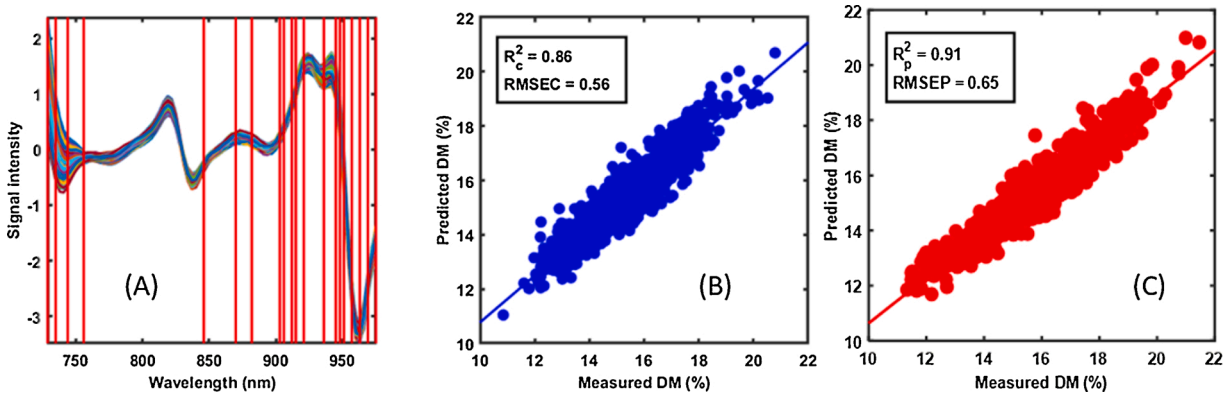

Fig. 5. Results of covariate selection (CovSel) and calibration for dry matter (DM %) prediction in apples. (A). Selected variables (vertical red lines), (B). calibration set, and (C). test set. R^2c: Coefficient of determination for calibration set, R^2p: coefficient of determination for test set, RMSEC: root mean squared error of calibration, and RMSEP: root mean squared error of prediction (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

which are related to the 3rd overtones of OH bonds, indicative of the moisture in the fresh fruit (Walsh et al., 2020b). Fig. 5B shows the results of calibration using the 20 selected variables and Fig. 5C shows the predictions of the calibrated model on the test set. The results indicate that CovSel variable selection reduced the RMSEP by 18 %, respectively, compared to the standard PLS regression modelling, suggesting that variable selection could be an approach to optimize prediction models. The results were in accordance with a recent study related to prediction of DM and SSC in new batch of pear fruit, where the model based on selected variable outperformed the standard PLS regression by attaining low RMSEP (Mishra et al., 2021). Variable selection maintains the model generalizability by retaining the key variables that are highly correlated with the property of interest; in other words, variable selection simplifies the models by leaving out the variables not necessarily related to the property of interest (Mehmood et al., 2012, 2020).

3.3. Model maintenance with dynamic orthogonal projections

The results of PLS cross-validation after the DOP correction are shown in Fig. 6. Both the explained variance and MSE were plotted as the function of the number of LVs to select the optimal complexity for calibrating the final model (Fig. 6). Compared to the cross-validation of the standard PLS regression (Fig. 3), DOP correction before the PLS regression gave an MSE inflexion point at same number of LVs, i.e. 6. A reason for this is the removal of the non-relevant information (external influences) by the DOP step prior to PLS calibration. The results of PLS modeling on DOP-corrected data, presented in Fig. 7, show that the RMSEP was reduced by 29 %, compared to standard PLS regression, suggesting that DOP may be highly beneficial for correcting unwanted effects in multi-batch data. In comparison to the CovSel variable selection approach, the DOP step attained 12 % lower RMSEP. The improvements with the DOP approach were in accordance with a recent study related to the use of DOP for improved prediction of DM in mango and olive fruit under various external effects such as temperature, instrument change and seasonal differences correction (Mishra et al., 2020b). However, a major limitation of DOP approach in comparison to variable selection and domain adaption techniques is that DOP requires new sample measurements to estimate the external influences in order to remove them (Zeaiter et al., 2006).

3.4. Domain adaption with transfer component analysis

The results of PLS regression modeling on TCA-corrected data (Fig. 8) show that the RMSEP reduced by 16 %, respectively, compared to standard PLS regression. In comparison to PLS regression, TCA seems to be effective to correct for unwanted variability in multi-batch data when no reference measurements are available from the new batch. The improvements with the TCA approach were in accordance with a recent study related to the use of TCA for improved prediction of DM in mango and olive fruit under various external effects such as temperature, instrument change and seasonal differences correction (Mishra et al., 2020b). With respect to DOP correction, the performances of TCA were worse in terms of RMSEP; however, TCA leads to a higher R^2 and it required only 3 LVs compared to the 5 required by DOP. In comparison to the variable selection with CovSel, TCA performed worse.

4. Conclusion

In the present work, a GUI has been presented for chemometric modelling of NIR spectra of fruit. Specific algorithms for correcting the batch effect were integrated. The GUI was demonstrated with a real-world dataset related to the prediction of DM in apple fruit. The results showed that the GUI can perform tasks such as PLS regression, variable selection and batch effect correction using either DOP or TCA methods. The results reported indicate that approaches such as variable selection and batch effect corrections with DOP and TCA can improve the performance of models built on NIR spectra collected in multi-batch experiments. The best performance in terms of lowest RMSEP was obtained with the DOP approach, followed by variable selection with CovSel and then TCA. However, a main drawback of the DOP approach is that it requires reference measurements from the new batch to model and remove the external influence. On the other hand, the main benefit of variable selection and TCA approaches is that they work without the need of any new reference measurements. In a practical scenario, it is recommended that the user exploit this GUI to compare multiple approaches to model NIR data, and, eventually, decide on the best approach for their specific challenges. The analysis presented in this article can be replicated by following the steps illustrated in the paper. The use of the GUI is not limited just to fruit but can be extended to all cases where the chemometric modelling methods presented here are required.

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Data curation, Investigation. **Jean Michel Roger:** Software, Visualization, Writing - review & editing. **Federico Marini:** Formal analysis, Software, Visualization, Writing - review & editing. **Alessandra Biancolillo:** Software, Visualization. **Douglas N. Rutledge:** Writing - review & editing.

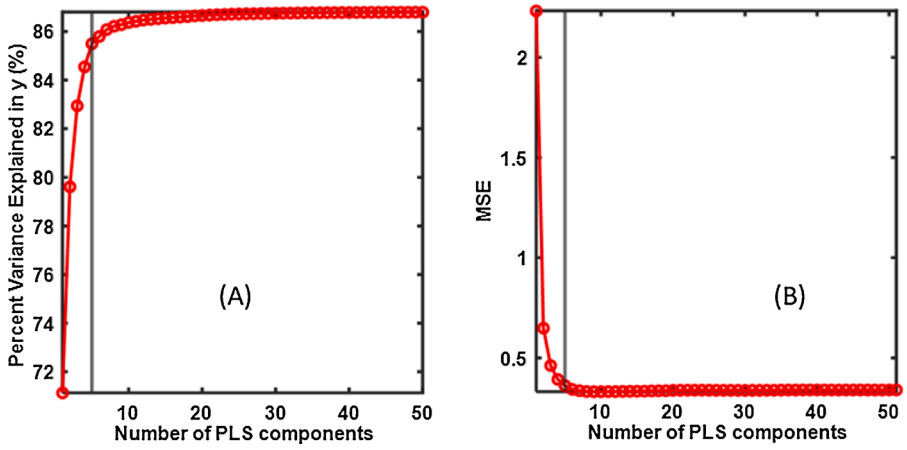

Fig. 6. Latent variables (LVs) optimization for partial least-squares (PLS) regression after dynamic orthogonal projections (DOP) correction. (A). Explained variance in response variables as a function of the number of LVs, and (B). Mean squared error (MSE: the mean of the squared residuals, i.e., differences between measured and predicted response) as function of the number of LVs.

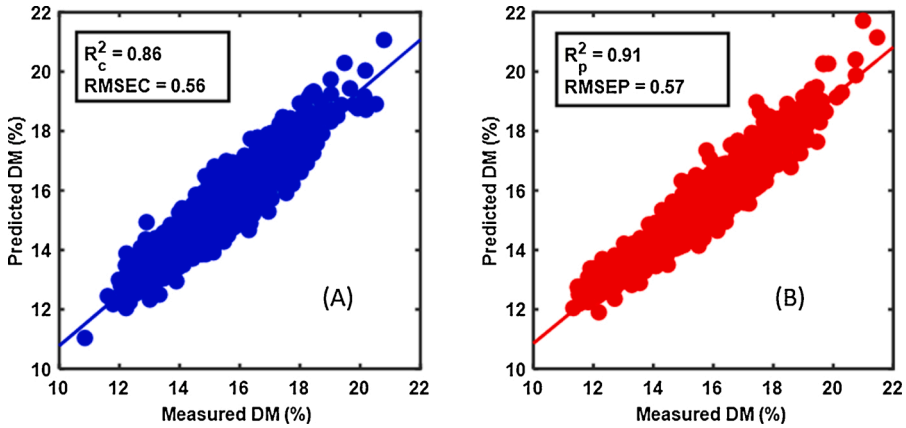

Fig. 7. Partial least-squares (PLS) regression calibration and testing after dynamic orthogonal projection (DOP) correction. (A) Calibration set, and (B) test set. R^2_c: Coefficient of determination for calibration set, R^2_p: coefficient of determination for test set, RMSEC: root mean squared error of calibration, and RMSEP: root mean squared error of prediction.

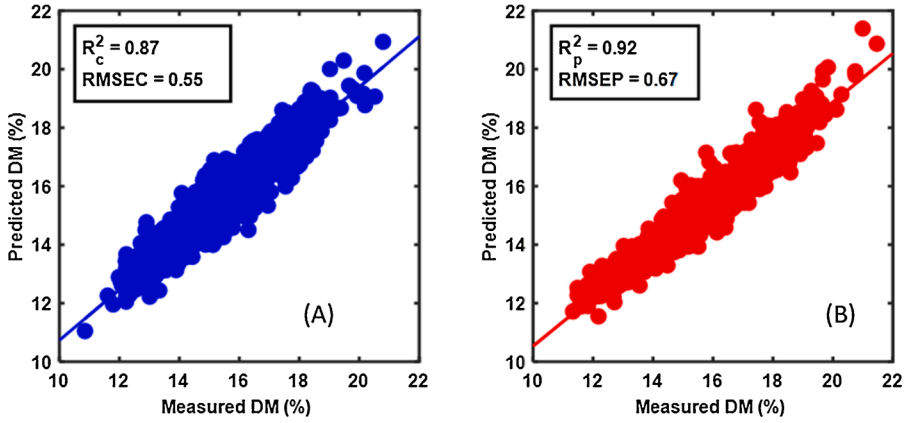

Fig. 8. Partial least-squares (PLS) regression calibration and testing after transfer component analysis (TCA) correction. (A) calibration set, and (B) test set. R^2_c: Coefficient of determination for calibration set, R^2_p: coefficient of determination for test set, RMSEC: root mean squared error of calibration, and RMSEP: root mean squared error of prediction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Dr. Soon Li The and Prof. Kate Evans from Washington State University for sharing the apple fruit multi-season dataset used for software demonstration.

References

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.

Boelens, H.F.M., Dijkstra, R.J., Eilers, P.H.C., Fitzpatrick, F., Westerhuis, J.A., 2004. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. J. Chromatogr. A 1057 (1), 21–30. https://doi.org/10.1016/j.chroma.2004.09.035.

Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., Walczak, B., 2007. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. Chemom. Intell. Lab. Syst. 85 (2), 269–277. https://doi.org/10.1016/j.chemolab.2006.03.006.

Dong, J.Y., Cheng, K.K., Xu, J.J., Chen, Z., Griffin, J.L., 2011. Group aggregating normalization method for the preprocessing of NMR-based metabolomic data. Chemom. Intell. Lab. Syst. 108 (2), 123–132. https://doi.org/10.1016/j.chemolab.2011.06.002.

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

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. https://doi.org/10.1016/S0003-2670(98)00737-5.

Isaksson, T., Næs, T., 1988. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Appl. Spectrosc. 42 (7), 1273–1284.

Mehmood, T., Liland, K.H., Snipen, L., Sæbø, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemom. 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. Chemom. e3226. https://doi.org/10.1002/cem.3226 n/a(n/a).

Mishra, P., Nikzad-Langerodi, R., 2020. Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit. Infrared Phys. Technol. 103547 https://doi.org/10.1016/j.infrared.2020.103547.

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., Lohumi, S., Ahmad Khan, H., Nordon, A., 2020a. Close-range hyperspectral imaging of whole plants for digital phenotyping: recent applications and illumination correction approaches. Comput. Electron. Agric. 178, 105780 https://doi.org/10.1016/j.compag.2020.105780.

Mishra, P., Roger, J.M., Rutledge, D.N., Woltering, E., 2020b. Two standard-free approaches to correct for external influences on near-infrared spectra to make models widely applicable. Postharvest Biol. Technol. 170, 111326 https://doi.org/10.1016/j.postharvbio.2020.111326.

Mishra, P., Roger, J.-M., Rutledge, D.N., Biancolillo, A., Marini, F., Nordon, A., Jouan-Rimbaud-Bouveresse, D., 2020c. MBA-GUI: a chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing. Chemom. Intell. Lab. Syst. 205, 104139 https://doi.org/10.1016/j.chemolab.2020.104139.

Mishra, P., Roger, J.M., Rutledge, D.N., Woltering, E., 2020d. 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.

Mishra, P., Biancolillo, A., Roger, J.-M., Marini, F., Rutledge, D.N., 2020e. New data preprocessing trends based on ensemble of multiple preprocessing techniques. Trac Trends Anal. Chem. 132, 116045 https://doi.org/10.1016/j.trac.2020.116045.

Mishra, P., Woltering, E., Brouwer, B., Hogeveen-van Echtelt, E., 2021. Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach. Postharvest Biol. Technol. 171, 111348 https://doi.org/10.1016/j.postharvbio.2020.111348.

Mobaraki, N., Amigo, J.M., 2018. HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis. Chemom. Intell. Lab. Syst. 172, 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003.

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.

Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2011. Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. 22 (2), 199–210. https://doi.org/10.1109/tnn.2010.2091281.

Rabatel, G., Marini, F., Walczak, B., Roger, J.-M., 2020. VSN: variable sorting for normalization. J. Chemom. 34 (2), e3164 https://doi.org/10.1002/cem.3164.

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. Chemom. Intell. Lab. Syst. 106 (2), 216–223. https://doi.org/10.1016/j.chemolab.2010.10.003.

Roger, J.-M., Boulet, J.-C., Zeaiter, M., Rutledge, D.N., 2020. Pre-processing methods. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Elsevier, pp. 1–75. https://doi.org/10.1016/B978-0-12-409547-2.14878-4.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36 (8), 1627–1639. https://doi.org/10.1021/ac60214a047.

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 pre-treatment. Postharvest Biol. Technol. 163, 111140 https://doi.org/10.1016/j.postharvbio.2020.111140.

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.

Walsh, K.B., Blasco, J., Zude-Sasse, M., Sun, X., 2020a. Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: the science behind three decades of commercial use. Postharvest Biol. Technol. 168, 111246 https://doi.org/10.1016/j.postharvbio.2020.111246.

Walsh, K.B., McGlone, V.A., Han, D.H., 2020b. 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.

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

Yan, Ke, 2020. A Domain Adaptation Toolbox (https://github.com/viggin/domain-adaptation-toolbox), GitHub. Retrieved November 4, 2020.

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. Chemom. Intell. Lab. Syst. 80 (2), 227–235. https://doi.org/10.1016/j.chemolab.2005.06.011.