Scientific African 3 (2019) e00051

Contents lists available at ScienceDirect

Scientific African

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

# Non-destructive measurement of internal quality of apple fruit by a contactless NIR spectrometer with genetic algorithm model optimization

Jean Frederic Isingizwe Nturambirwe[a,b], Helene H. Nieuwoudt[c], Willem J. Perold[a], Umezuruike Linus Opara[b,*]

[a] Department of Electrical and Electronic Engineering, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
[b] Postharvest Technology Research Laboratory, South African Research Chair in Postharvest Technology, Department of Horticultural Science, Stellenbosch University, Private Bag X1, Stellenbosch 7602, South Africa
[c] Institute for Wine Biotechnology, Department of Viticulture and Oenology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

ARTICLE INFO

*Article history:*
Received 5 November 2018
Revised 1 February 2019
Accepted 2 February 2019

*Keywords:*
FT-NIR
Calibration transfer
Model performance
Genetic algorithm
Apple quality

ABSTRACT

Spectrometric methods based on near infrared radiation (NIR) are commonly used effectively in the agricultural and food industry. However, these methods still face limitations whereby meeting requirements for application such as nondestructive quality testing of large fruits and automated sorting and grading is still a challenge. A Fourier transform (FT)-NIR spectrometer (emission head, EH mode of Matrix-F) that simulates on-line sample scanning (contactless, large sample size (100 mm)) was used to predict internal properties of apple fruit. The EH was compared to laboratory multipurpose analyzer (MPA) FT-NIR spectrometer using two contact-sample presentation modes with relatively smaller sample size ($\leq$22 mm); namely, the integrating sphere (IS) and the solid probe (SP). Three apple cultivars (Golden Delicious, Granny Smith and Royal Gala) sourced from two retail stores (in Stellenbosch, South Africa) were used to constitute variability in the sample set. Partial least squares regression (PLSR) prediction models for internal quality (total soluble solids (TSS) and titratable acidity (TA)) were developed and validated on external test samples in various scenarios. Genetic algorithm (GA) based optimization of PLS models was used to produce optimal models prior to instrumental comparison.

Model optimization using GA improved performance by a margin of 30% of the original root mean square error of cross validation for the contactless system bringing it closer to the performance of models from the MPA. The EH's performance makes it an attractive

*Abbreviations:* EH, emission head; MPA, Multipurpose analyzer; RG, Royal gala; SD, standard deviation; Ch, Checkers retail store; 1der, First derivative; M-Mnorm, Min-Max normalization; SP, solid probe; GD, Golden delicious; cal, Calibration; correl, correlation; FLM, Food lovers' market; SLS, Straight line subtraction; SNV, Single normal variate; IS, Integrating sphere; GS, Granny Smith; val, Validation; var, Variables; LV, Latent variables; MSC, Multiple scatter correction.

* Corresponding author.
*E-mail address:* opara@sun.ac.za (U.L. Opara).

https://doi.org/10.1016/j.sciaf.2019.e00051
2468-2276/© 2019 The Authors. Published by Elsevier B.V. on behalf of African Institute of Mathematical Sciences / Next Einstein Initiative. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

option for achieving on-line application of NIR spectroscopy for sorting apples based on internal quality.

© 2019 The Authors. Published by Elsevier B.V. on behalf of African Institute of Mathematical Sciences / Next Einstein Initiative.
This is an open access article under the CC BY license.
(http://creativecommons.org/licenses/by/4.0/)

## Introduction

Spectroscopic methods have gained increasing interest in quality evaluation of food commodities. Spectral data are transformed into useful information by means of chemometrics. The latter combines multivariate statistical analysis and spectral processing methods to establish relationships between quantifiable quality attributes and spectral data. The targeted uses for NIR spectroscopy application in horticultural industry include mainly fruit sorting and grading based on internal quality attributes [30]. Such grading is important because internal attributes such as sugar content and acidity, among others, contribute considerably in ensuring consumer satisfaction as well as meeting requirements for specific protocols in food processing. The performance of calibration models developed from these spectroscopic methods and the effectiveness of statistical methods used can be limited by the experimental conditions such as spectral acquisition system accuracy or precision, the nature of spectral data (highly correlated variables), etc. Since the NIR spectra contains a lot of irrelevant information, given a specific problem, selection of variables with chemical relevance is often required [29,31]. Also, variable selection outcome can help determine the most relevant filters for the application of NIR on-line [28,31].

Various numerical methods for variable selection and optimization have been used in combination with multivariate statistical analysis in order to improve prediction models [28]. Artificial neural networks were used in combination with genetic algorithms (GA-ANNs) for the nondestructive quantitative analysis of cefalexin based on NIR reflectance spectra [8]. Fei et al. [8] reportedly conveyed that since GA is a global search method, it has less probability to be trapped at local minima; its combination with ANNs (implemented as the fitness function for GA) would perform better than many other selection methods. Their study showed that GA improved the performance of ANNs, which, on the other hand, proved to give better models than PLS [8]. Genetic algorithms were used with multivariate regression to determine gelatine in historic paper using infrared and NIR data. The model obtained using GA was built on fewer data points (76 vs. 2150) and latent variables (4 vs 9) than that based on full spectra [6]. More information on variable selection and optimization techniques as summarized in Xiaobo et al. [28] shows possibilities to improve prediction models and there is still room for improvement.

One of the hurdles that hinder widespread use of NIR systems is that of calibration transfer. A calibration model developed on one instrument may not be directly usable on another, even if they are from the same manufacturer or are the same model. Having to construct the calibration model for every spectrometer is expensive and time consuming. These difficulties are associated to changes in the instrument response due to aging or maintenance and environmental factors such as temperature and humidity variations [7,10]. The latter can have a strong influence on measurement values by causing shifts in absorption bands and non-linear changes in absorption intensities, among other factors [27]. Other than 'standardization' methods that have proven to be useful in addressing the calibration transferability [1,12]; in case of absorbance shift related problems and when instrumental differences are small, there are alternative approaches to solve the transfer problem. Some of these approaches include using appropriate pre-processing methods, wavelength selection and including data acquired by several instruments in the calibration [9,23,24].

In this work two NIR sample exposure modes, namely the integrating sphere, IS and solid probe, SP of the multipurpose analyzer NIR spectrometer (Bruker Optics, Germany) were used as reference in order to assess their similarities and or divergences from the matrix-F (Bruker Optics, Germany) in emission head (EH) mode. The kind of similarities that could work in favor of alleviating the issue of calibration transfer between these spectrometers. The objectives were 1) to assess possible differences and/or similarities between the reference spectrometer modes and the emission head based on regression model predictive ability and spectral profile; 2) to apply wavelength selection and pre-treatment methods that are appropriate for model simplification in predicting internal attributes of apple fruit and 3) derive insights on the issues of calibration transferability there associated.

## Material and methods

### *Sampling*

Apples were purchased in two installments (in two consecutive months) from two different retail shops in Stellenbosch, South Africa. A batch of 100 apples were sourced first (source denoted Ch in Fig. 1) and 114 apples were acquired in the second instance (source denoted by FLM in Fig. 1). Three cultivars of apple, namely Golden Delicious (yellowish green), Granny Smith (green) and Royal Gala (predominantly red), were acquired in both instances, in nearly equal proportions (see Table 1). The fruits were kept in cold storage (5 °C) pending Fourier transform (FT)-NIR spectral acquisition and

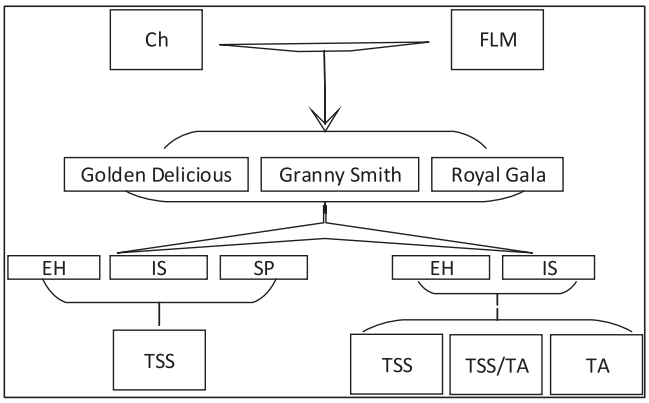


**Fig. 1.** Experimental setup for NIR measurements of internal attributes on apples. A summary of data acquisition systems, cultivars and respective targeted attributes for the analysis. FLM, source 2; Ch, source 1; EH, 'emission head'; IS, 'integrating sphere'; SP, 'solid probe'; TSS, total soluble solids; TA, titratable acidity; TSS/TA, sugar:acid ratio.

**Table 1**
An overview of the reference measurements for internal quality attributes.

| Source | Instrument | Attribute | Cultivar | N | Mean | SD | Range |
|---|---|---|---|---|---|---|---|
| | EH / IS | TA (%) | All | 228 | 2.68 | 0.29 | 2.09 – 3.6 |
| FLM | | TSS (°Brix) | " | " | 13.41 | 1.92 | 9 – 18.3 |
| | | TSS/TA | " | " | 5.09 | 1.02 | 2.5 – 7.5 |
| | EH / IS | | GD | 66 | 13.38 | 1.23 | 11.1 – 16.1 |
| | | TSS (°Brix) | GS | 62 | 12.64 | 1.41 | 8.8 – 15.8 |
| | | | RG | 72 | 15.09 | 1.28 | 12.8 – 17.7 |
| | | | All | 200 | 13.77 | 1.67 | 8.8 – 17.7 |
| Ch | | | | | | | |
| | SP | | GD | 32 | 13.12 | 1.14 | 11.1 – 14.9 |
| | | TSS (°Brix) | GS | 32 | 12.09 | 1.38 | 8.8 – 15.1 |
| | | | RG | 32 | 14.97 | 1.16 | 12.9 – 17.3 |
| | | | All | 96 | 13.39 | 1.72 | 8.8 – 17.3 |
| All | | | | | 13.54 | 1.43 | 8.8 – 18.3 |

FLM, fruit material source 2; Ch, fruit material source 1; EH, 'Emission head' of the Matrix-F; IS, 'integrating sphere' of the MPA; SP, 'solid probe' of the MPA; TSS, total soluble solids; TA, titratable acidity; TSS/TA, sugar:acid ratio; GD, Golden delicious; GS, Granny Smith; RG, Royal Gala; SD, standard deviation; Range, min – max of measured values; Instrument, sample exposure mode used for respective sample batches; N, number of samples.

destructive measurements thereafter. They were left at room temperature for three hours to equilibrate prior to experiments. Fig. 1 summarizes the experimental design from sampling to NIR spectrometer modes and destructively targeted attributes. All three spectrometer modes were used to scan fruit from source Ch and only TSS was destructively measured whereas, only EH and IS were used for spectral acquisition and both TSS and TA were measured on samples from FLM.

*Destructive measurements*

Both non-destructive and destructive measurements were carried out on whole fruit samples. Destructive measurements involved the measurement of total soluble solids and titratable acids. Total soluble solids (TSS) content was measured by slicing a small portion of apple tissue from both sides of the apple where the NIR spectra were acquired, and squeezing out the sliced tissue's juice on the lens of a hand-held digital refractometer (Palette, PR-32 a, Brix 0.0–32.0, Atago Co. Ltd., Japan) for reading and expressed in °Brix. A refractometer calibration with distilled water was required before commencing the actual measurement for every sample. Titratable acidity (TA), on the other hand, was measured on apple juice from blended fruits, whereby individual samples were prepared from single apple juice separately. TA values were acquired by titrating 2 mL of juice against 0.1 N NaOH to an end point at pH = 8.2 using a compact titrosampler (862 Compact Titrosampler®, Metrohm, Switzerland) and the '2 mL' juice method, and expressed in% of juice. The values for total tiratable acids were used for the calculation of a third, derived attribute, that represents the sugar:acid ratio (TSS/TA).

*NIR spectroscopy measurements*

Non-destructive measurements were performed by means of near infrared spectroscopic techniques. Two opposite points around the equatorial plane of every apple were scanned on two different spectrometers of which three different sample

exposure modes were used, namely the solid probe and integrating sphere modes of the multipurpose analyzer (MPA; Bruker Optics, Germany), and the non-contact emission head of the Matrix-F spectrometer (Matrix-F duplex from Bruker Optics, Germany). For each single measurement the spectrum was averaged over 64 scans. The NIR scanning range was between 12,500 – 4000 $cm^{-1}$, in intervals of 4 $cm^{-1}$ [20]. A brief overview of the data acquisition in relation to the quality attributes targeted in the data analysis is given by the chart in Fig. 1.

The solid probe uses a permanently aligned and highly stable Rock Solid$^{TM}$ interferometer and a 20 W Tungsten halogen lamp as NIR source. The interferometer is equipped with high reflective surface and inert, gold coated mirrors and has a wavenumber accuracy and precision better than 0.1 $cm^{-1}$ and 0.04 $cm^{-1}$ respectively. The beam splitter is made of a quartz substrate with proprietary coating. The position and velocity of the movable mirror is accurately calculated using a He-Ne class 1 laser. The fiber optic probe contains both, in a bifurcated optical configuration, the source fibers that guide the light to the sample, which is in direct contact with the optic probe, and the detector fibers that receive the reflected light [3].

The integrating sphere mode is used to measure diffuse reflectance of highly scattering solid media. It is associated with 50 mm width sample cup holder (22 mm spot size) for measurements of heterogeneous samples, on which apples were placed for scanning. The integrating sphere uses the same spectroscopic elements as for the fiber optic probe channel, except for the detector. The integrating sphere makes use of a high sensitivity PbS detector with non-linearity correction. An internal gold reference spectrum was obtained by mechanically closing the optical window with a gold reference plate.

The MATRIX-F FT-NIR spectrometer is equipped with a fiber optic NIR illumination and detection head (185 mm height and 230 mm diameter for sample sizes up to 100 mm in diameter) and allowed for measurement on almost half of the fruit surface in a single exposure . The fiber optic illumination head contains 4 air cooled tungsten NIR light sources (Tungsten halogen, 12 V, 20 W). The diffusely reflected light from the sample is collected and guided via a fiber optic cable to the spectrometer detector (a highly sensitive, thermoelectric cooled and temperature controlled InGaAs diode detector) [4].

*Data analysis*

Multivariate data analysis methods were used to explore spectral data and to build and optimize prediction models of quality attributes. The methods comprised of principal component analysis, partial least squares (PLS) regression and genetic algorithm (GA) coupled with PLS, which was used for variable selection. Principal component analysis (PCA) is a technique for reducing the amount of data when there is correlation present, which is common in NIR data. It is worth stressing that it is not a useful technique if the variables are uncorrelated [22]. It approximates a data matrix, X (N objects $\times$ K variables), by the product of two matrices T and P' that capture the essential data patterns of X. By so doing, a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables [14]. Many goals can be achieved through PCA including simplification, data reduction, modeling, outlier detection, classification, variable selection, etc. [25]. PCA was used in this work mainly to explore the variability in the data and classification.

PLS regression (PLSR) is a method for relating two data matrices, X (predictors) and Y (response), by a linear multivariate model, but goes beyond traditional regression in that it models also the structure of X and Y [22]. Its ability to analyze data with many, noisy, collinear, and even missing data in both X and Y makes it very useful. PLSR has the desirable property that the precision of the model parameters improves with the increasing number of relevant variables and observations [26]. Many PLS algorithms have been developed, including the orthogonal score PLS, on which most variable selection methods are based [21]. In this work, PLS regression methods were used to establish models for predicting internal quality of apples. Spectral data were averaged per fruit (two spectra per sample) and mean-centered. Prediction models were built using 25% of the samples set as validation set. For every y variable (TSS, TA, and TSS/TA) an individual model was established. Prediction models were built in different scenarios that take into account variabilities such as the effect of cultivar, sample source and spectrometer mode on model performance. The models robustness was investigated by following a validation procedure shown by the descriptive chart in Fig. 2.

OPUS software version 7.2 (from Bruker Optik GmbH) was used for spectral acquisition and processing thereafter, 64 scans were averaged to make up a single spectrum. PLS regression analysis subsequent to spectral preprocessing was also performed in OPUS software, which has a feature of selective search based on model performance associated to preprocessing method or methods combination. The best performance as rated by the software in search for best preprocessing method was based upon the lowest value of the root mean square error of prediction (RMSEP).

Furthermore, a genetic algorithm described by Leardi [16] was applied to the dataset in an attempt to improve prediction model performance. The algorithm starts by creating a population of randomly structured and individually unique chromosomes made of binary encoded variables (1 for selected variables and 0 for excluded ones). First, PLS regression is applied to each subset and the measure of best model performance is used to determine the fittest chromosomes. In each iteration, the GA applies crossover and mutation operators to the existing population to create a new population of subsets (offspring). The crossover randomly reorders one pair from the solution. Then, it iteratively exchanges elements in any position of the two subsets with a probability $p_c$. Mutation iteratively modifies each element within the subset with a probability $p_m$. The new population is updated by selecting the best chromosomes among the offspring and the parent chromosomes (current population) altogether. The process goes on until the stop criterion (predefined number of evaluations) is reached [6,15,32].

The PLS-GA models were evaluated according to the values of the RMSECV and the coefficient of determination $R^2$ i.e. better models would have lowest RMSECV and highest $R^2$ values.

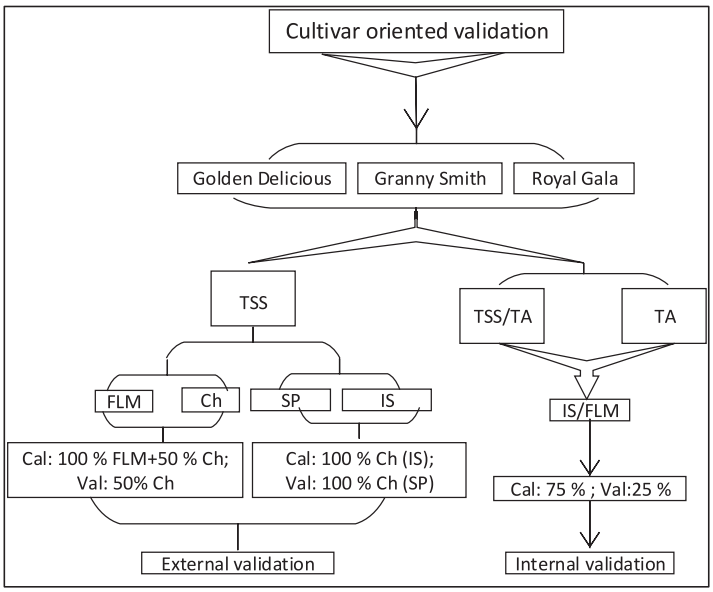


**Fig. 2.** Model validation summary applied to three cultivars altogether. FLM, source 2; Ch, source 1; EH, 'Emission head'; IS, 'integrating sphere'; SP, 'solid probe'; TSS, total soluble solids; TA, titratable acidity; TSS/TA, sugar:acid ratio; GD, Golden Delicious apple cultivar; GS, Granny Smith apple cultivar; RG, Royal Gala apple cultivar; Cal, calibration; Val, validation.

The parameters of the GA were as follows [15]:

- population size: 30 chromosomes;
- regression method: PLS;
- response to maximize: cross-validated explained variance;
- leave-out groups: 5;
- average number of variables per initial chromosome: 5;
- $p_c$: 0.5;
- $p_m$: 0.01;
- maximum number of PLS components: 15;
- number of runs: 100;
- the amount of evaluations: 200.

## Results and discussion

### *Spectral analysis*

NIR spectra of all apple cultivars had a similar profile in all acquisition modes with four main peaks around 10,270, 8620, 6890 and 5200 $cm^{-1}$. Fig. 3 compares the baseline corrected spectra of six random apples from three different spectrometer modes: Red (MPA - SP), green (Matrix-F - EH) and blue (MPA - IS). The spectral peaks around 5200 $cm^{-1}$ (1923 nm) and at 6890 $cm^{-1}$ (1449 nm) were relatively higher from the solid probe than for spectra from both the IS and the EH, while the peaks at 10,269 $cm^{-1}$ and 8623 $cm^{-1}$ were generally lowest for the solid probe.

These differences could be related to differences in fruit surface area scanned specific for each acquisition mode. The peaks around 10,270 and 6890 $cm^{-1}$ corresponded to the 2nd and 1st vibrational overtones of OH stretching associated with water absorption [2,5]. On the other hand, the peaks around 8620 and 5200 $cm^{-1}$ correspond to the 2nd and 1st overtones of CH stretching, as well as the 3rd overtone of OH, CH and $CH_2$ deformation associated with sugar solution [19].

### *Data distribution*

Fruit were sourced from two different supermarkets in different monthly periods, spectral data acquired in different spectral acquisition modes and three reference quality parameters were measured destructively. Table 1 gives a brief overview of the quantitative measurements done on the entire dataset in different categories.

Measures of soluble solids ranged from 8.8 to 18.3 $^{o}$Brix, titratable acidity from 2.09 to 3.6%, both resulting in values of sugar:acid ratio ranging from 2.5 to 7.5. In the batch of fruits bought from Checkers, the measures of TSS were also shown for different cultivars separately. Royal Gala apples had the highest mean value of TSS followed by Golden Delicious apples.

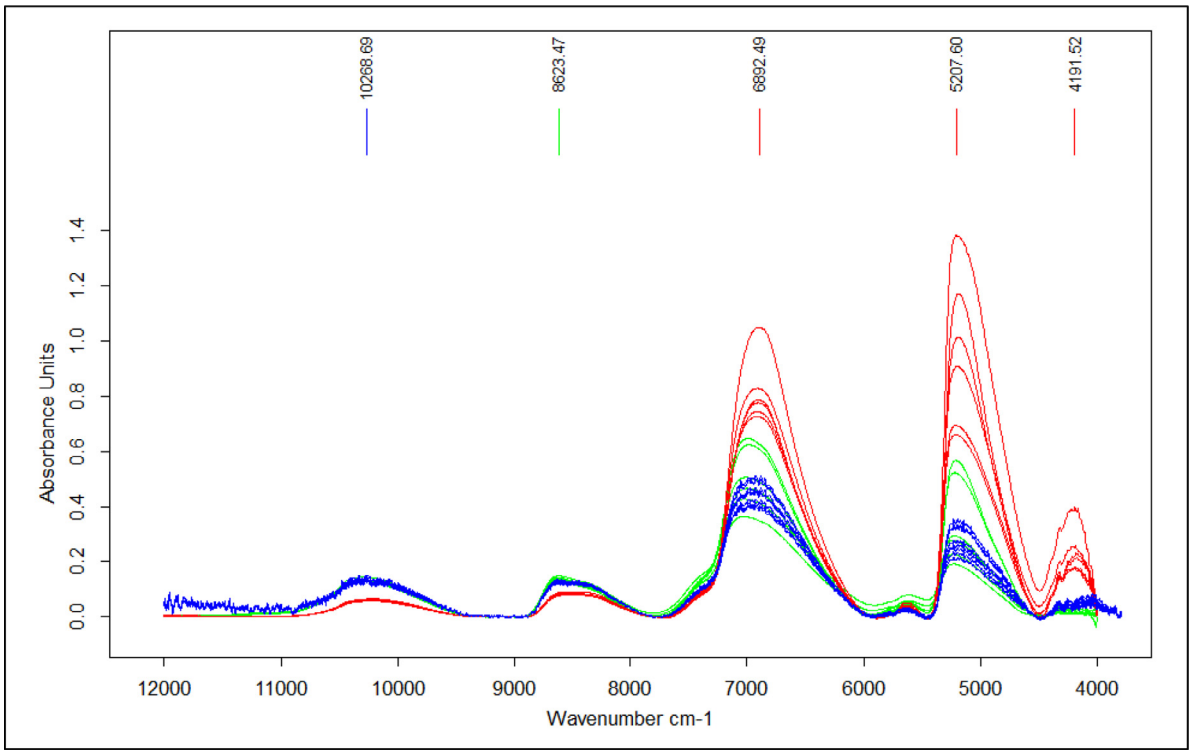


**Fig. 3.** Comparison of spectra in three acquisition modes: spectra from solid probe, red; for integrating sphere, blue and from emission head, green.

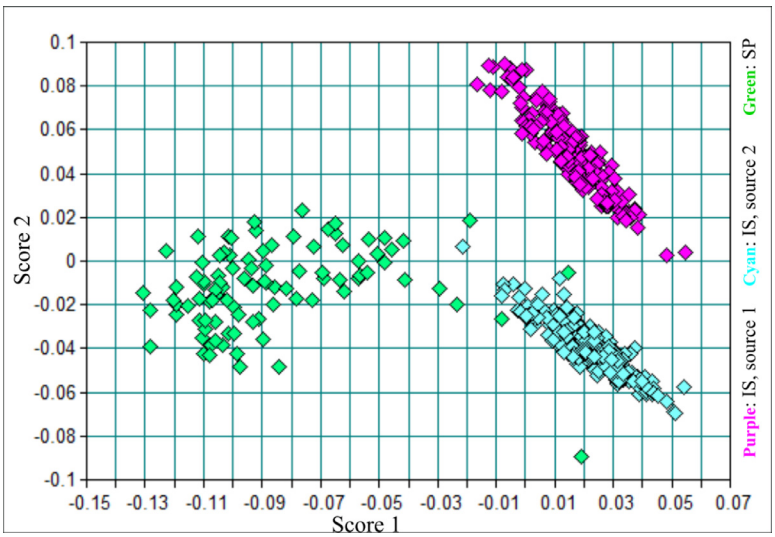


**Fig. 4.** PCA scores plot for all spectral data acquired on the MPA. The first principal component separates samples with respect to spectrometer modes (green for 'MPA - SP' versus the rest for 'MPA - IS'). The second component clearly separates sample sources 'Ch' (purple) versus 'FLM' (Cyan).

The reference values were normally distributed around the respective mean values and over a range that is large enough to constitute a good dataset for meaningful data analysis. TSS was spread over a range of 9.5 ºBrix, which is more than half the maximum of TSS values. A similar measured range was found in the values for the other reference quality attributes.

Spectral data exploration using principal component analysis (PCA) resulted in clusters of samples differentiated with respect to spectrometer mode and sample sources (Fig. 4). The first principal component separated data with respect to sample acquisition modes (SP versus IS) from the MPA. The second component helped distinguish between sample sources whereby samples from 'Ch' were clearly separated from those from 'FLM', all acquired from the IS.

Similar classes were obtained with the first two components, using most of the preprocessing methods [18], except for the first derivative and straight line subtraction (SLS) methods, where the third principal component distinguished best between sample sources instead of the second. These results highlight the variability in the dataset used here and suggest its relevance for the purpose of external data validation introduced in Section "Data Analysis" and the validation procedure shown in Fig. 2.

*Total soluble solids*

The measured soluble solids content in apples were used as reference values in building prediction models by NIR spectroscopy in all the three exposure modes (IS, SP and EH). Two spectra acquired per apple, each from either side, were averaged as well as the TSS values from both scanned sides and used as a single sample. A prediction model was built based on full spectral data and improved by means of spectral pre-processing. The best pre-processing method was chosen according to the performance of the resulting prediction model and constituted the subject of the report presented here. The model performance was rated based on parameters such as coefficient of determination, $R^2$; the relative prediction deviation, RPD; the error of prediction, RMSEP/CV; latent variables, LV and the slope. The best performing pre-processing methods for TSS differed from those obtained in TA and TSS/TA, and were also different with respect to sample sources. SLS (straight line subtraction) was the best pre-processing method for TSS in samples from 'Ch', whilst for 'FLM' samples SNV (standard normal variate) led to the best prediction model parameters (RMSEP, $R^2$, RPD and slope). Fruit samples from 'Ch' were used in models predicting total soluble solids (TSS) only. The prediction set was generated by selecting a block of two out of five consecutive samples up to a number that makes up to about 25% of all the samples; the remaining samples, approximately 75%, were used for model estimation. External validation of samples from IS for TSS used samples acquired from the MPA probe, and vice versa. Fruits from 'FLM' were used to build models based on TSS, TA and TSS/TA, using both the EH and IS acquisition modes. While SNV was the best performing preprocessing method for TSS and dominant in all the three attributes, SNV combined with first derivative was dominant in models involving TA (TA and TSS/TA).

The predictive model for TSS with SNV as the pre-processing method of spectra acquired on the EH for all the three apple cultivars had the coefficient of determination, $R^2$, varying between 89.09% and 96.73% and the RMSEP between 0.365 and 0.40.

*Titratable acidity*

Predicting titratable acids was best achieved by using the first derivative as pre-processing method. The best model was obtained within restricted wavelength regions (9403.5–7498.1 $cm^{-1}$; 6101.9–5774 $cm^{-1}$). Unlike TSS, TA models were very mediocre ($R^2 < 50\%$) without any spectral pre-processing. A tremendous improvement in the prediction model was achieved after pre-processing, resulting in the value of $R^2 = 68.17\%$, with a very low error of prediction of RMSEP = 0.12.

*Soluble solids to titratable acids ratio (TSS/TA)*

TSS/TA is commonly used as a good indicator of maturity in various types of fruit, including apples. The pre-processing method that resulted in the best prediction model for TSS/TA was the combination of both the best pre-processing methods for TSS and TA, i.e. 1st derivative and SNV. A good prediction model was obtained for TSS/TA, with $R^2 = 82.62\%$, RMSEP = 0.43, based on data from the Matrix-F (EH) and models based on other instruments were summarized in Table 3. TSS models differed slightly with respect to the type of validation set (sample source or exposure mode). It was noticed that lower RMSEP and higher RPD and $R^2$ values were obtained for samples from FLM, where the test set was from the same source and acquisition mode, than in the case where the test set was external (different source or exposure mode, see Table 3). Therefore, external data validation induced more variability and slightly reduced the model performance, but as commonly understood, such a validation contributes to model robustness [11,20].

Prediction models for TA had the lowest performance (lowest coefficients of determination, lowest RPD, slope farthest from 1, largest difference between prediction and calibration $R^2$ values) relatively to TSS and TSS/TA models. The EH had better predictive ability than the IS. In all cases, the three internal quality indicators in apple fruit (TSS, TA and TSS/TA) were well predicted by means of FT-NIR in different modes of acquisition.

*Effect of cultivar on prediction models*

Three apple cultivars were used in this work, namely Golden Delicious (GD), Granny Smith (GS) and Royal Gala (RG). There have been more studies on apple quality focusing on single cultivar than those combining many cultivars at a time [13,29]. A combined study provides a way to highlight the effect of biological variability, if any, for a specific investigation, given that there is always differences from one cultivar to another.

The values summarized in Table 2 are based on NIR spectral data that were acquired using the MPA (both IS and SP). Internal validation (25% test vs 75% calibration) was used on samples from 'FLM', whilst external validation was performed on the batch 'Ch' by choosing samples from the SP as test set and those from the IS as calibration set (see Fig. 2). Best prediction models for TSS were found in GD followed by GS and then RG, for samples from 'FLM' (with internal validation),

**Table 2**
A summary of FT-NIR prediction model parameters as related to apple cultivar.

| Source | Attribute | Cult | LV | Preproc | Calibration | | | Validation | | | | Waveband ($cm^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $R^2$ (%) | RMSEC | Slope | $R^2$ (%) | RMSEP | RPD | Slope | |
| | | GD | 8 | None | 97.08 | 0.32 | 0.971 | 94.34 | 0.333 | 4.22 | 0.925 | 9403.7–6098.1/5450.1–4246.7 |
| | TSS (°Brix) | GS | 8 | M-Mnorm | 96.95 | 0.271 | 0.969 | 91.48 | 0.374 | 3.44 | 0.903 | 9403.7–5446.3 |
| | | RG | 8 | M-Mnorm | 95.91 | 0.305 | 0.959 | 86.49 | 0.483 | 2.73 | 0.834 | 6102–4246.7 |
| . | | | | | | | | | | | | |
| | | GD | 6 | 1der+MSC | 73.44 | 0.077 | 0.734 | 50.01 | 0.133 | 1.52 | 0.417 | 9403.7–8451/5176.3–4246.7 |
| FLM | TA (%) | GS | 6 | None | 54.09 | 0.123 | 0.541 | 31.21 | 0.196 | 1.35 | 0.359 | 5450.1–4597.7 |
| | | RG | 7 | 1der+MSC | 76.58 | 0.084 | 0.766 | 69.1 | 0.125 | 1.8 | 0.573 | 6102–5446.3/4601.6–4246.7 |
| . | | | | | | | | | | | | |
| | | GD | 9 | M-Mnorm | 97.51 | 0.139 | 0.975 | 91.19 | 0.251 | 3.48 | 0.83 | 9403.7–7498.3/5450.1–4246.7 |
| | TSS/TA | GS | 8 | None | 81.79 | 0.212 | 0.818 | 71.48 | 0.225 | 1.95 | 0.795 | 9403.7–8451/5450.1–5022 |
| | | RG | 7 | SLS | 84.9 | 0.22 | 0.849 | 71.94 | 0.327 | 1.89 | 0.592 | 9403.7–7498.3/6102–5446.3 |
| . | | | | | | | | | | | | |
| | | GD | 5 | 1der+MSC | 95.37 | 0.388 | 0.954 | 75.81 | 0.545 | 2.1 | 0.652 | 9403.7–5446.3 |
| Ch | TSS (°Brix) | GS | 10 | SLS | 98.63 | 0.178 | 0.986 | 80.72 | 0.608 | 2.3 | 0.747 | 9403.7–6098 |
| | | RG | 9 | 1der+SLS | 95.61 | 0.328 | 0.956 | 85.61 | 0.442 | 2.64 | 0.85 | 9403.7–7498.3/6102–5446.3 |

FLM, source 2; Ch, source 1; TSS, total soluble solids; TA, titratable acidity; TSS/TA, sugar:acid ratio; Cult, cultivar; GD, Golden delicious apple cultivar; GS, Granny Smith apple cultivar; RG, Royal Gala apple cultivar; RMSEC, root mean square error of calibration; RMSEP, root mean square error of prediction; LV, latent variables; Preproc, pre-processing method; RPD, relative prediction deviation; SLS, straight line subtraction;.M-Mnorm, min-max normalization; MSC, multiplicative scatter correction; 1der, first derivative; None, no spectral pre-processing.

while this order was reversed in samples from 'Ch' (with external validation). TA was best predicted in RG apples followed by GD and then in GS apples. A similar order to that in TSS (GD > GS > RG), but in a different scenario, was found in the prediction model for TSS/TA. The best model was obtained in GD, while the model parameters in the remaining cultivars were not outstandingly distinct. We argue that, even though the $R^2$ value in RG was slightly higher than that found for GS, it was noticed that GS had the lowest error of prediction of TSS/TA, a better RPD and Slope than the one obtained for RG. The difference between $R^2$ values for calibration and validation was also lower in GS than in RG. Therefore, it was concluded that the predictive model for GS was better than that for RG.

*Comparison of spectrometer modes*

It has not been possible to develop an 'all purpose' FT-NIR system, even though multiple functions or uses may be performed on the same system. It is understood that spectrometers designed differently are also likely to perform differently when used for the same tasks. However, there are ways of circumventing such hurdles with model optimization. For example, the EH of the Matrix-F used in this project was designed for process monitoring and allows for much larger sample sizes. The MPA on the other hand, although equipped with multiple modes of sample exposure, has limitations when it comes to large (>5 cm) sample sizes. The MPA can however be instrumental in comparing measurements from different designs of the same analytical method for validation purposes and the development of new models. Here, comparison of optimized prediction models that were built based on data from three different sample exposure modes and two different FT-NIR spectrometers was carried out.

The prediction model parameters summarized in Table 3 are a comparative overview of the spectrometer modes (EH of the Matrix-F spectrometer; IS and SP modes of the MPA) used in this work. The relative predictive ability of the spectrometers was dependent on quality parameters. The SP mode had the lowest predictive ability for TSS ($R^2$ = 0.82, highest RMSEP = 0.57 °Brix and lowest RPD = 2.73) relatively to the EH ($R^2$ = 0.88, RMSEP = 0.51 °Brix, RPD = 2.9) and the IS ($R^2$ = 0.90, RMSEP = 0.57 °Brix, RPD = 3.24) with comparable slopes, where external validation was used. Nonetheless, the SP did give the lowest number of latent variables and the highest slope, which contributes to a relatively simpler model. In the case where internal validation was used, the IS performed consistently better (higher $R^2$ and slope, better RPD and lower RMSEP) than the EH in predicting both TSS and TSS/TA. However, the predictive parameters were very close (only different to the hundredth) in value and with the same optimal wavebands and pre-processing method, in the case of TSS. The IS and the EH displayed a near identical ability to predict TSS. On the other hand, the EH outperformed the IS in the measurements for predicting TA.

*Application of GA-PLS for internal quality prediction*

A genetic algorithm designed to optimize PLS regression models [17] was used in order to study the improvement of model performance in different scenarios (two different spectral acquisition systems, three different apple cultivars) and for effective wavelength selection. For a typical prediction model based on spectra acquired with the EH, a plot of predicted versus true values of TSS is shown in Fig. 5. Although, the prediction performance indicators ($R^2$, RMSEP, RPD) can be rated as good, there is a considerable margin (i.e., from $R^2$ = 90.63–100%) of needed improvement for more accuracy. Genetic algorithm based optimization of PLS models was performed to this end.

**Table 3**
A summary of prediction models for internal attributes. The '*' and '**' indicate where external validation based on acquisition mode and source were used, respectively.

| Source | Attrib | Instr | LV | Preproc | Calibration | | | Validation | | | | Waveband ($cm^{-1}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $R^2$ (%) | RMSEC | Slope | $R^2$ (%) | RMSEP | RPD | Slope | |
| CH | | EH | 6 | SLS | 90.47 | 0.53 | 0.91 | 89.05 | 0.49 | 3.09 | 0.87 | 9403.5–7498.1; 4601.5–4424.1 |
| * | TSS | IS | 10 | COE | 94.71 | 0.38 | 0.95 | 90.48 | 0.57 | 3.24 | 0.81 | 9403.7–7498.3; 6102–5446.3 |
| * | | SP | 5 | COE | 85.66 | 0.67 | 0.86 | 81.87 | 0.57 | 2.73 | 0.93 | 9403.7–7498.3; 4601.6–4246.7 |
| ** | | EH | 10 | 1der | 92.86 | 0.49 | 0.93 | 87.93 | 0.51 | 2.9 | 0.87 | 7502–5446.2 |
| | TSS | EH | 10 | SNV | 97.14 | 0.32 | 0.97 | 97.1 | 0.35 | 5.97 | 0.95 | 9403.7–5446.2; 4601.6–4246.7 |
| | | IS | 10 | SNV | 97.97 | 0.27 | 0.98 | 97.21 | 0.32 | 5.99 | 0.98 | 9403.7–5446.2; 4601.6–4246.7 |
| FLM | | | | | | | | | | | | |
| | TA | EH | 6 | 1der+SNV | 72.11 | 0.16 | 0.72 | 68.17 | 0.18 | 1.79 | 0.65 | 9403.5–7498.1; 6101.9–5774 |
| | | IS | 7 | SLS | 75.46 | 0.15 | 0.76 | 58.62 | 0.19 | 1.57 | 0.60 | 9403.5–7498.1; 6102–5446.3 |
| | TSS/TA | EH | 6 | 1der+SNV | 86.83 | 0.37 | 0.87 | 82.62 | 0.43 | 2.4 | 0.86 | 9403.5–7498.1; 6101.9–5446.2 |
| | | IS | 9 | SNV | 91.73 | 0.30 | 0.92 | 91.57 | 0.28 | 3.75 | 1.05 | 7425–5446.3; 4601.6–4424.1 |

FLM, fruit source 2; Ch, fruit source 1; EH, 'Emission head'; IS, 'integrating sphere'; SP, 'solid probe'; TSS, total soluble solids; TA, titratable acidity; TSS/TA, sugar:acid ratio; COE, constant offset elimination; SNV, vector normalization; RMSEC, root mean square error of calibration; RMSEP, root mean square error of prediction; LV, latent variables; Preproc, pre-processing method; RPD, relative prediction deviation; SLS, straight line subtraction.

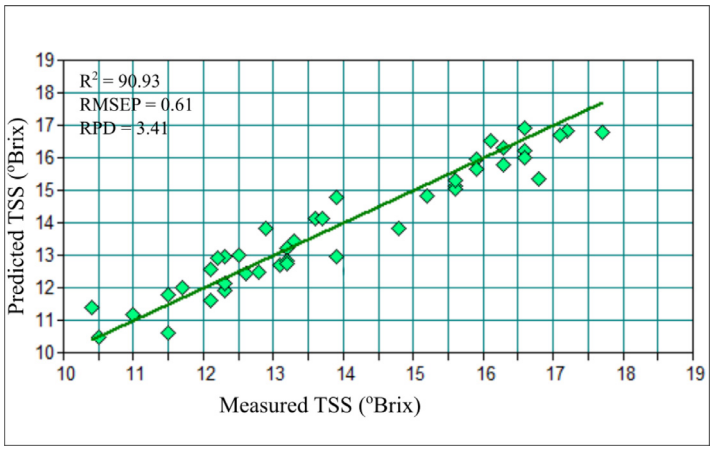


**Fig. 5.** Prediction of total soluble solids based on spectra from EH. $R^2$, coefficient of determination; RMSEP, root mean square error of prediction; RPD, relative predictive deviation.

First, GA-PLS was performed on full spectra, and variable selection was evaluated in comparison to previous research findings [31]. Average contiguous spectral data were also used as an extended tool to confirm the accuracy of the free full spectra-based GA run. Risks of over fitting may be encountered when the number of variables largely exceeds that of the observations, which is very likely to happen for spectral data like in NIR spectroscopy. In order to check for the possibility that the GA calculations could have been impaired by the so-called 'large p problem' (for a "$n \times p$" data matrix X), the average contiguous wavelengths were used to reduce the spectral data to less than 200 variables. Every 11 consecutive variables were averaged and used as 1, reducing the full spectra from 2074 to 188 variables for the IS (MPA) and from 2307 to 192 variables (12 variables averaged to 1) for the EH (Matrix). The results of variable selection in both cases led to model performances that were closely similar (Table 4). The values in Table 4 give a comparative overview of model performance when GA was (sections 'GA-PLS' and 'Avcont') or wasn't (section 'PLSR') applied to PLS regression. These values were averaged over five individual runs and standard deviations are indicated. For the PLS latent variables (LV), the statistical mode across the five GA runs was indicated instead. The PLSR models were built using 10-fold cross validation, without any pre-processing methods, but outliers were deleted from the models.

The % explained variance expressed by the coefficient of determination ($R^2$) in cross-validation was relatively higher in all GA models than the full spectra PLSR models. The number of variables in the GA models was reduced by more than a factor of 12 relative to the full spectra models. GA did not reduce the number of latent variables, but remained comparable in the same attributes. The error of cross-validation expressed as RMSECV was improved (reduced by 30% for the EH and by 24% for the IS) by GA in models for TSS, but remained relatively the same in models for TA. The performance of models that were built based on average contiguous variables was relatively the same as those from GA applied on the original variables. The

**Table 4**
Full-spectrum PLS and GA optimized PLS model performance for predicting soluble solids and titratable acidity in apples.

| | $R^2$ (%) | SD | Var | SD | LV | SD | RMSECV | SD | Attribute | |
|---|---|---|---|---|---|---|---|---|---|---|
| EH | 96.16 | 0.11 | 150.80 | 2411 | 11.00 | 0.98 | 0.38 | 0.01 | TSS | GA-PLS |
| IS | 97.12 | 0.09 | 148.60 | 22.90 | 11.00 | 0.80 | 0.32 | 0.01 | | |
| EH | 54.01 | 0.30 | 99.20 | 37.59 | 7.00 | 0.40 | 0.20 | 0.00 | TA | |
| IS | 59.78 | 1.73 | 106.80 | 42.50 | 7.00 | 0.00 | 0.19 | 0.00 | | |
| EH | 95.69 | 0.10 | 61.80 | 9.52 | 12.00 | 0.80 | 0.40 | 0.00 | TSS | Avcont |
| IS | 96.86 | 0.04 | 85.40 | 17.67 | 12.00 | 0.98 | 0.34 | 0.00 | | |
| | | Slope | Correl | Bias | | RPD | | | | |
| EH | 91.23 | 0.91 | 0.96 | −0.02 | 10.00 | 3.38 | 0.57 | | TSS | PLSR |
| IS | 95.27 | 0.94 | 0.98 | −0.02 | 10.00 | 4.60 | 0.42 | | | |
| EH | 42.06 | 0.46 | 0.65 | 0.00 | 6.00 | 1.31 | 0.21 | | TA | |
| IS | 54.07 | 0.58 | 0.74 | 0.00 | 9.00 | 1.48 | 0.19 | | | |

'GA-PLS', Genetic Algorithm with Partial Least Squares regression; 'Avcont', Average contiguous wavelengths were used in the GA-PLS model development; 'SD', standard deviation of the adjacent parameter (in table) for five runs; 'Correl', correlation coefficient between predicted and real values; 'Var', number of variables used in the final model; 'LV', latent variables; 'RPD', relative prediction deviation; 'EH', emission head mode; 'IS', Integrating sphere mode; 'RMSECV', root mean square error of cross validation; 'TSS', total soluble solids; 'TA', titratable acids.

IS consistently provided better models than those based on the EH for both modeling approaches and both quality attributes. Nonetheless, both instruments displayed a relatively close performance in predicting these attributes (Table 4).

It should be noted that the relative performance of the acquisition modes discussed here was also realized for external validation involving optimization based on pre-processing of spectra reported in Table 3.

## Conclusions

This work reported on predicting internal quality of apple fruit non-destructively, using three sample exposure configurations from two FT-NIR spectrometers. The main objective was to evaluate the fitness of the EH of the Matrix-F spectrometer, designed for online applications, in assessing apple quality. The EH's performance was compared to the common laboratory MPA as our reference performance standard.

NIR spectral data were used to build models to predict some indicators of apple internal quality (TSS, TA and TSS/TA). Various case scenarios were used to assess the performance of models, namely the effect of cultivar and spectrometer modes on the models' performance. The models were also optimized by using different preprocessing techniques in various wavelength regions of the entire spectra and genetic algorithm.

Results suggested that there were differences in spectral intensities with respect to spectrometer modes, but the same spectral profile. Relative prediction performances with respect to cultivars, per single attribute, varied depending on factors such as validation approach and spectrometer mode. Genetic algorithm improved the performance of EH by a larger margin that the IS, resulting in close indicator values of their prediction performances.

The comparison of spectrometer modes revealed that, although the IS seemed to outperform the EH in predicting TSS and TSS/TA and the opposite in predicting TA, the model parameters were close in value in most of the cases and both modes performed relatively better than the SP of the MPA. Similar results were obtained in models optimized using genetic algorithm. The EH configuration, given its capability for online sample scanning and its demonstrated performance in this work, is therefore a fit system for rapid assessment of internal quality of apple fruit and therefore a prominent candidate for industrial application.

In light of the demonstrated similarities in performance of the EH of the Matrix-F and the IS mode of the MPA, it is likely that the transferability of calibrations from the IS mode to the EH mode would present less challenges than usually encountered in this subject matter.

## Declarations of interest

None.

## Acknowledgments

The authors are thankful to Ricardo Leardi for freely providing the software tools to implement genetic algorithm in this project.

## Funding

This work is based on the research supported wholly by the National Research Foundation of South Africa (Grant Numbers: 64813). The opinions, findings and conclusions or recommendations expressed are those of the authors alone, and the NRF accepts no liability whatsoever in this regard.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sciaf.2019.e00051.

## References

[1] E.L. Bergman, H. Brage, M. Josefson, O. Svensson, A. Sparén, Transfer of NIR calibrations for pharmaceutical formulations between different instruments, J. Pharm. Biomed. Anal. 41 (2006) 89–98. https://doi.org/10.1016/j.jpba.2005.10.042.
[2] E. Bobelyn, A.S. Serban, M. Nicu, J. Lammertyn, B.M. Nicolai, W. Saeys, Postharvest quality of apple predicted by NIR-spectroscopy: study of the effect of biological variability on spectra and model performance, Postharvest Biol. Technol. 55 (2010) 133–143. https://doi.org/10.1016/j.postharvbio.2009.09.006.
[3] Bruker, MPA II Multi Purpose FT-NIR Analyzer [WWW Document], 2018. URL https://www.bruker.com/products/infrared-near-infrared-and-raman-spectroscopy/ft-nir/mpa/overview.html.
[4] Bruker, n.d. Matric-F FT-NIR Spectrometer [WWW Document]. URL https://www.bruker.com/products/infrared-near-infrared-and-raman-spectroscopy/ft-nir/matrix-f/overview.html.
[5] C. Camps, P. Guillermin, J.C. Mauget, D. Bertrand, Discrimination of storage duration of apples stored in a cooled room and shelf-life by visible-near infrared spectroscopy, J. Near Infrared Spectrosc. 15 (2007) 169–177. https://doi.org/10.1255/jnirs.726.
[6] L. Cséfalvayová, M. Pelikan, I. Kralj Cigić, J. Kolar, M. Strli, Use of genetic algorithms with multivariate regression for determination of gelatine in historic papers based on FT-IR and NIR spectral data, Talanta 82 (2010) 1784–1790. https://doi.org/10.1016/j.talanta.2010.07.062.
[7] T. Fearn, Standardisation and calibration transfer for near infrared instruments: a review, J. Near Infrared Spectrosc 9 (2001) 229–244.
[8] Q. Fei, M. Li, B. Wang, Y. Huan, G. Feng, Y. Ren, Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection, Chemom. Intell. Lab. Syst. 97 (2009) 127–131. https://doi.org/10.1016/j.chemolab.2009.03.003.
[9] R.N. Feudale, H. Tan, S.D. Brown, 2002. Piecewise orthogonal signal correction 63, 129–138.
[10] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Transfer of multivariate calibration models: a review, Chemom. Intell. Lab. Syst. 64 (2002) 181–192.
[11] M. Golic, K.B. Walsh, Robustness of calibration models based on near infrared spectroscopy for the in-line grading of stonefruit for total soluble solids content, Anal. Chim. Acta 555 (2006) 286–291. https://doi.org/10.1016/j.aca.2005.09.014.
[12] C.V. Greensill, K.B. Walsh, Calibration transfer between miniature photodiode array-based spectrometers in the near infrared assessment of mandarin soluble solids content, J. Near Infrared Spectrosc. 10 (2002) 27–36.
[13] P. Jannok, S. Kawano, Development of a common calibration model for determining the Brix value of intact apple, pear and persimmon fruits by near infrared spectroscopy, J. Near Infrared Spectrosc. 22 (2014) 367. https://doi.org/10.1255/jnirs.1130.
[14] I.T. Jolliffe, Principal component analysis, Encyclopedia of Statistics in Behavioral Science, 2nd ed., 2002 https://doi.org/10.2307/1270093.
[15] R. Leardi, A.L. Gonzalez - genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207, doi:10.1016/S0169-7439(98)00051-3.
[16] R. Leardi, Application of genetic algorithm–PLS for feature selection in spectral data sets, J. Chemom. 14 (2000) 643–655.
[17] R. Leardi, L. Nörgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, 2005. https://doi.org/10.1002/cem.893.
[18] L.C. Lee, C. Liong, A.A. Jemain, A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, Chemom. Intell. Lab. Syst. 163 (2017) 64–75. https://doi.org/10.1016/j.chemolab.2017.02.008.
[19] L.S. Magwaza, Non-destructive Prediction and Monitoring of Postharvest Quality of Citrus Fruit, Stellenbosch University, 2013.
[20] L.S. Magwaza, U.L. Opara, L.A. Terry, S. Landahl, P.J.R. Cronje, H.H. Nieuwoudt, A. Hanssens, W. Saeys, B.M. Nicolaï, Evaluation of Fourier transform-NIR spectroscopy for integrated external and internal quality assessment of Valencia oranges, J. Food Compos. Anal. 31 (2013) 144–154. https://doi.org/10.1016/j.jfca.2013.05.007.
[21] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.
[22] J.M. Miller, J.C. Miller, Statistics and chemometrics for analytical chemistry, technometrics, 2010. https://doi.org/10.1198/tech.2004.s248.
[23] H. Swierenga, P.J. de Groot, A.P. de Weijer, M.W.J. Derksen, Improvement of PLS model transferability by robust wavelength selection, Chemom. Intell. Lab. Syst. 41 (1998) 237–248.
[24] P. Tillmann, T.-C. Reinhardt, C. Paul, Networking of near infrared spectroscopy instruments for rapeseed analysis: a comparison of different procedures, J. Near Infrared Spectrosc. 8 (2000) 101–108.
[25] S. Wold, K.I.M. Esbensen, P. Geladi, Principal component analysis. Chemom, Intell. Lab. Syst. 2 (1987) 37–52.
[26] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.
[27] F. Wulfert, W.T. Kok, O.E. de Noord, A.K. Smilde, Correction of temperature-induced spectral variation by continuous piecewise direct standardization, Anal. Chem. 72 (2000) 1639–1644. https://doi.org/10.1021/ac9906835.
[28] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32. https://doi.org/10.1016/j.aca.2010.03.048.
[29] Z. Xiaobo, Z. Jiewen, H. Xingyi, L. Yanxiao, 2007. Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji ' apple based on different PLS models 87, 43–51. https://doi.org/10.1016/j.chemolab.2006.09.003.
[30] H. Xu, B. Qi, T. Sun, X. Fu, Y. Ying, Variable selection in visible and near-infrared spectra: application to on-line determination of sugar content in pears, J. Food Eng. 109 (2012a) 142–147. https://doi.org/10.1016/j.jfoodeng.2011.09.022.
[31] H. Xu, B. Qi, T. Sun, X. Fu, Y. Ying, Variable selection in visible and near-infrared spectra: application to on-line determination of sugar content in pears, J. Food Eng. 109 (2012b) 142–147. https://doi.org/10.1016/j.jfoodeng.2011.09.022.
[32] Y.H. Yun, D.S. Cao, M.L. Tan, J. Yan, D.B. Ren, Q.S. Xu, L. Yu, Y.Z. Liang, A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration, Chemom. Intell. Lab. Syst. 130 (2014) 76–83. https://doi.org/10.1016/j.chemolab.2013.09.007.