Research Article

A sampling approach for predicting the eating quality of apples using visible–near infrared spectroscopy

Mabel V Martínez Vega,[a]* **Sara Sharifzadeh,**[b] **Dvoralai Wulfsohn,**[c] **Thomas Skov,**[d] **Line Harder Clemmensen**[b] **and Torben B Toldam-Andersen**[a]

Abstract

BACKGROUND: Visible–near infrared spectroscopy remains a method of increasing interest as a fast alternative for the evaluation of fruit quality. The success of the method is assumed to be achieved by using large sets of samples to produce robust calibration models. In this study we used representative samples of an early and a late season apple cultivar to evaluate model robustness (in terms of prediction ability and error) on the soluble solids content (SSC) and acidity prediction, in the wavelength range 400–1100 nm.

RESULTS: A total of 196 middle–early season and 219 late season apples (*Malus domestica* Borkh.) cvs 'Aroma' and 'Holsteiner Cox' samples were used to construct spectral models for SSC and acidity. Partial least squares (PLS), ridge regression (RR) and elastic net (EN) models were used to build prediction models. Furthermore, we compared three sub-sample arrangements for forming training and test sets ('smooth fractionator', by date of measurement after harvest and random). Using the 'smooth fractionator' sampling method, fewer spectral bands (26) and elastic net resulted in improved performance for SSC models of 'Aroma' apples, with a coefficient of variation CV_{SSC} = 13%. The model showed consistently low errors and bias (PLS/EN: R^2_{cal} = 0.60/0.60; SEC = 0.88/0.88°Brix; $Bias_{cal}$ = 0.00/0.00; R^2_{val} = 0.33/0.44; SEP = 1.14/1.03; $Bias_{val}$ = 0.04/0.03). However, the prediction acidity and for SSC (CV = 5%) of the late cultivar 'Holsteiner Cox' produced inferior results as compared with 'Aroma'.

CONCLUSION: It was possible to construct local SSC and acidity calibration models for early season apple cultivars with CVs of SSC and acidity around 10%. The overall model performance of these data sets also depend on the proper selection of training and test sets. The 'smooth fractionator' protocol provided an objective method for obtaining training and test sets that capture the existing variability of the fruit samples for construction of visible–NIR prediction models. The implication is that by using such 'efficient' sampling methods for obtaining an initial sample of fruit that represents the variability of the population and for sub-sampling to form training and test sets it should be possible to use relatively small sample sizes to develop spectral predictions of fruit quality. Using feature selection and elastic net appears to improve the SSC model performance in terms of R^2, RMSECV and RMSEP for 'Aroma' apples.
ⓒ 2013 Society of Chemical Industry

Keywords: *Malus domestica*; SSC; representative sample; training set formation; variability

INTRODUCTION

The use of visible and near infrared spectroscopy (visible–NIR) for the rapid evaluation of fruit quality remains a topic of importance and interest for the food research community and food industry because, in a near future, it might be included in 'the tool box' for efficient farm management.[1,2] Spectral regions on the visible and near infrared spectrum have been used to predict quality in intact fruits such as apples (380 up to 2000 nm), apricots (600–2500 nm), citrus (636–1236 nm), grapes (650–1100 nm), kiwifruits (300–1100 nm), pineapples (400–2500 nm) with different degrees of success.[3] The fruit quality parameters studied with spectroscopy included: soluble solids content (SSC), firmness, acidity, dry matter, taste and starch, among others.[3] In most of these studies the quality characteristics were predicted using multivariate statistical models.

Two of the most important fruit quality traits are SSC and acidity.[4] These traits have a great influence on consumer liking

and repetitive purchases. During fruit growth, the internal quality traits are expected to vary due to different causes (type of soil, weather, training and thinning techniques, etc.). This variation in quality might be the most important factor affecting the

* Correspondence to: *Mabel V Martínez Vega, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Højbakkegård Allé 13, 2630 Taastrup, Denmark. E-mail: mmar@life.ku.dk*

a *Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Højbakkegård Allé 13, 2630 Taastrup, Denmark*

b *DTU Data Analysis, The Technical University of Denmark, Richard Petersens Plads, build. 305/123, 2800 Lyngby, Denmark*

c *Dayenú Ltda, San Fernando, Chile*

d *Quality and Technology, Department of Food Sciences, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C., Denmark*

calibration models, which are used to train different spectroscopy devices.[3] Model validation, an essential step to be carried out after calibration, has often been performed using samples from the same batch. The tendency has been to use or suggest large sets of samples, which together with pre-processing statistical methods, reached somewhat satisfactory results.[3−5] One conclusion was that the samples should be 'rich' in variation and ideally contain information from multiple orchards/seasons/cultivars to obtain sufficient robustness.[3,6−8] In addition, for the purpose of proper model construction, post-harvest sample arrangements have also been proposed with different aims. Interestingly, most of the studies reporting spectral robustness issues for fruit quality, used samples obtained randomly from either fruit trees or from the commercial market. Frequently, little has been reported regarding the sampling techniques applied during fruit collection and often relevant sample statistics (mean, standard deviation, ranges of the quality parameter of interest) have not been provided. As a result, the reproduction, comparison, evaluation and improvement of the mentioned experiments becomes challenging.

In an earlier study, we explored the variability of mass, sugar, firmness and starch of representative samples of 'Granny Smith' apples obtained at the orchard scale (Martínez Vega MV et al., unpublished). In this study, we extend our approach of using the 'fractionator' tree sampling procedure to obtain representative apple fruit samples at time of harvest.[9] These samples were used to evaluate the performance of visible–NIR spectroscopy method for calibration and validation model development. Thus, the objectives of the study were: (1) evaluate the SSC and acidity prediction performance of an early and late season apple cultivar; and (2) to compare different sub-sampling techniques to form training and test sets on the overall performance of the prediction models. Furthermore, we discuss the main implications of the method in practice.

MATERIALS AND METHODS

Fruit material

Two Danish apple (*Malus domestica* Borkh.) cultivars, an early season 'Aroma' and a late season 'Holsteiner Cox', were collected at fruit maturity, in September and October 2011, respectively from 11-year-old trees at the Pometum orchard, University of Copenhagen, Denmark. The samples were selected using the 'fractionator' procedure for trees,[9] from 10 trees per cultivar. The fractionator procedure is a form of multi-level systematic uniform random cluster sampling, in which the trees, primary branches, and, at the final stage, branch segments form the clusters of fruit for sampling purposes. For both cultivars we used systematic sampling periods of 2 (for branch) and 2 for in-branch segment with random starts. When the branch segments bore more than one fruit, a random number was used to select one fruit from each of the final sample of branch segments. Each sampled fruit was labelled with a number to preserve information about the picking order. Once harvested, the samples were kept at room temperature (18 °C). To widen the spread of fruit SSC and acidity values for the experiments, apple quality measurements for each cultivar were performed after 5 (Date 1) and 10 (Date 2) days of storage. Likewise, to preserve the distribution of fruits per tree from the original sample, the sub-groups for each of the storage periods mentioned were selected by taking a systematic sample of fruit with period 2 while preserving the original picking order from the 10 trees. Thus, sample Date 1 contained apples 1, 3, 5 . . . , and Date 2 sample consisted of the fruit ranked 2, 4, 6 and so forth.

Determination of fruit quality

On each apple, two pieces of fruit flesh (stem to calix end), from the exposed and non-exposed side of the fruit were squeezed. Its juice was presented to a calibrated handheld brix meter (Metler Toledo 'Quick brix 60'; Mettler Toledo Inc. Columbus, Ohio, USA) to measure SSC content. The remaining juice was kept for acidity determination. Acidity was measured with a titrino (719 S Titrino Metrohm; , Herisau, Switzerland). The titration consisted of adding a solution of NaOH of concentration 0.1 mol L^{-1} to 5 mL a sample solution of apple juice until the pH reached 8.1. Results were expressed in grams of malic acid (the most abundant acid in apples) per 100 mL of apple juice.

Spectral measurements

A spectrometer (MOE-1 System, Tec5 AG, Oberursel, Germany) with MMS sensors and a 12 V/100 W halogen lamp was used to collect reflectance readings in 1 nm increments within a wavelength range between 400–1130 nm, yielding 731 values per spectrum. A calibration was performed using a white piece of barium sulfate every 20 apples. Spectral measurements were performed on the exposed and non-exposed (to sun) parts around the equator of each apple. A distance between the lamp and the fruit of 10 mm was maintained. A holder supported fruits to direct light in a 45° angle to avoid specular reflectance. The integration time was 161 ms. Each intact fruit was placed on a rotary circular base with the stem–calyx vertical and four equidistant guides on the base made sure that the measurements were approximately equidistant. The scans collected at each sample point were averaged and transformed to absorbance [log(1/reflectance)].[10]

Training and test sets arrangements

First, over-mature or damaged fruit samples on 'Date 2' were removed from the data sets. Then, three different data sets were formed for each cultivar.

Set A

A smooth arrangement from all samples ('Date 1' and 'Date 2' together) according to SSC and acidity values was performed. The 'smooth' arrangement was formed by ranking all the original sample of fruit in increasing order according to the SSC or acidity level, respectively, for SSC and acidity modelling. Then every second fruit was pushed out to form a monotonically increasing and then decreasing ordering of fruit by quality. From this new ordering, a predefined systematic sampling interval of '4' (probability $p = 1/4$) was applied to obtain approximately 25% of the samples for the test set. The remaining 75% of the samples comprised the training set. This procedure was repeated four times, starting with fruit ranked 1, 2, 3 and 4, corresponding to the four possible 'random starts' that form all possible systematic samples from the 'smooth' ordering. Systematic sampling from a smooth arrangement ('smooth fractionator') is a procedure designed to provide samples with high within-sample variance and low between-sample variance, which in this case means that both training and test sets capture well the variation of SSC and of acidity existing in the original sample.[11,12] The averages of results were used to evaluate the general performance of the regression methods on the data sets.

Set B

SSC samples of each cultivar from 'Date 1' formed the training set and samples from 'Date 2' the test set. The same criterion was used to construct the models for acidity.

Set C

The original data set was divided into training (75%) and test (25%) sets using simple random sampling without replacement. This was repeated 25 times to obtain 25 independent sets for training and testing. The averages of results were used to evaluate the general performance of the regression methods on the data sets.

Preprocessing of spectral data

Since the spectral data contained NIR bands, Multiplicative scatter correction (MSC) was applied.[10] In addition, because of the presence of visual bands, the original data set without MSC was also considered. All the models and algorithms were calculated using Matlab software (version R2011a; The MathWorks Inc., Natick, MA, USA).

Calculation of calibration and prediction models

Three different linear regression methods were used on each data set. For all the regression methods, 10-fold cross validation with a modified version of the standard error rule[13] was used for finding the best parameters to train the model.

Partial least squares regression

The commonly used partial least squares regression (PLS) method was used to predict fruit quality from spectra data. The basis of the method is to link the variation in the spectral information to the response to find only the relevant information for predicting the response.[10]

Calibration and prediction models were constructed using the 'internal validation' approach (using samples from the same batch).[3] The SSC and acidity data were autoscaled before model calculation. This latter procedure ensures that all samples have approximately the same contribution to the model.[10]

Ridge regression

This method is based on the penalisation of the regression coefficients. As a result, the regression model is regularised to reduce the variance of the predicted output.[13] The purpose is to alleviate the effect of noise on the model. Ridge regression requires that both the response vector (Y) and the data matrix (X) to be centred.

Elastic net

Elastic net (EN) is a sparse regression method based on the regularisation of regression coefficients. This means that the regression coefficients are shrunk so that some of them are set to zero. Therefore, it can cancel out the noise effect. In addition, it has a grouping effect and the non-zero coefficients correspond to the groups of correlated variables (wavelengths). When the number of variables (e.g. number of spectral bands = 731) is higher than the number of observations (e.g. $N_{SSC} = 196$ data points), the prediction becomes an 'ill-posed' problem[13] and EN is one of the appropriate methods in this case. This method requires the response vector (Y) to be centred and the data matrix (X) to be normalised with unique length for each variable.[14]

Feature selection

This method is commonly used for high dimensional data to reduce the complexity of the model. Since the dimensionality of the apple data was high (731 spectral bands), this pre-processing

step was also employed. It was compared with the regression results using all the features (wavelengths). Feature selection helps to distinguish the wavelengths that carry the useful information for the prediction to simplify the model.

A common approach for dimension reduction is principal component analysis (PCA), but it is not an appropriate method for 'ill-posed' problems.[15] Although PCA is a dimension reduction method, each principal component is a linear combination of all the basic features (wavelengths). This means that it could not be used as a tool for reducing the number of used wavelengths for prediction. We applied a feature selection algorithm proposed in a former study (Sharifzadeh S *et al.*,[16] unpublished). The method first sorts the wavelengths according to the number of times that their corresponding regression coefficients were non-zero in several iterations of elastic net regression and then selects a subset of them as described below.

For this research, the regression coefficients obtained from applying EN on the set C (25 randomly generated training sets), were used for feature selection. First, the number of times that the coefficients were non-zero in each band was counted ('frequency of being non-zero'). Then, the wavelengths were sorted according to their corresponding frequencies. To choose a proper number of wavelengths for performing the regression task, a candidate list of the number of selected wavelengths was formed:

candidate list of top selected wavelengths

$$= [20, 50, 80, 100, 150, 200, 250, 700]$$

In the next step, an EN regression with 10-fold cross validation was applied on only the 25 training sets using the spectral data corresponding to each of these candidate numbers of wavelengths. Finally, the best candidate number of wavelengths was chosen according to the corresponding minimum root mean square error of prediction (RMSEP).

Model evaluation

Model robustness was evaluated in terms of the coefficient of determination (R^2), the standard deviation (SD) of training and test sets, the standard error of calibration (SEC), the root mean square of residual errors of cross validation (RMSECV), the standard error of prediction (SEP), the root mean square of residual errors of prediction (RMSEP) and the bias.

RESULTS AND DISCUSSION

Cultivar variability along the harvest season

The 'fractionator' procedure yielded in total 205 fruits for 'Aroma' and 221 fruits for 'Holsteiner Cox'. The total number of samples for 'Aroma' in Date 1 was $N = 103$ fruit and for Date 2 was $N = 102$ fruit. The number of samples for 'Holsteiner Cox' was $N = 111$ fruit in Date 1 and $N = 110$ fruit in Date 2. Figure 1 illustrates the spread of SSC and acidity values for both cultivars after elimination of damaged samples (over-mature or with disease). The higher SSC values of 'Holsteiner Cox' were expected given the reported sweetness properties of the late season cultivar as compared to the early season 'Aroma'.[17]

In general SSC and acidity values had low to moderate variation. 'Aroma' samples had the same average of SSC on both post-storage dates, whereas 'Holsteiner Cox' samples showed a slight increase of the average SSC values. The increase is related to degradation of starch which normally is present at high levels in late season
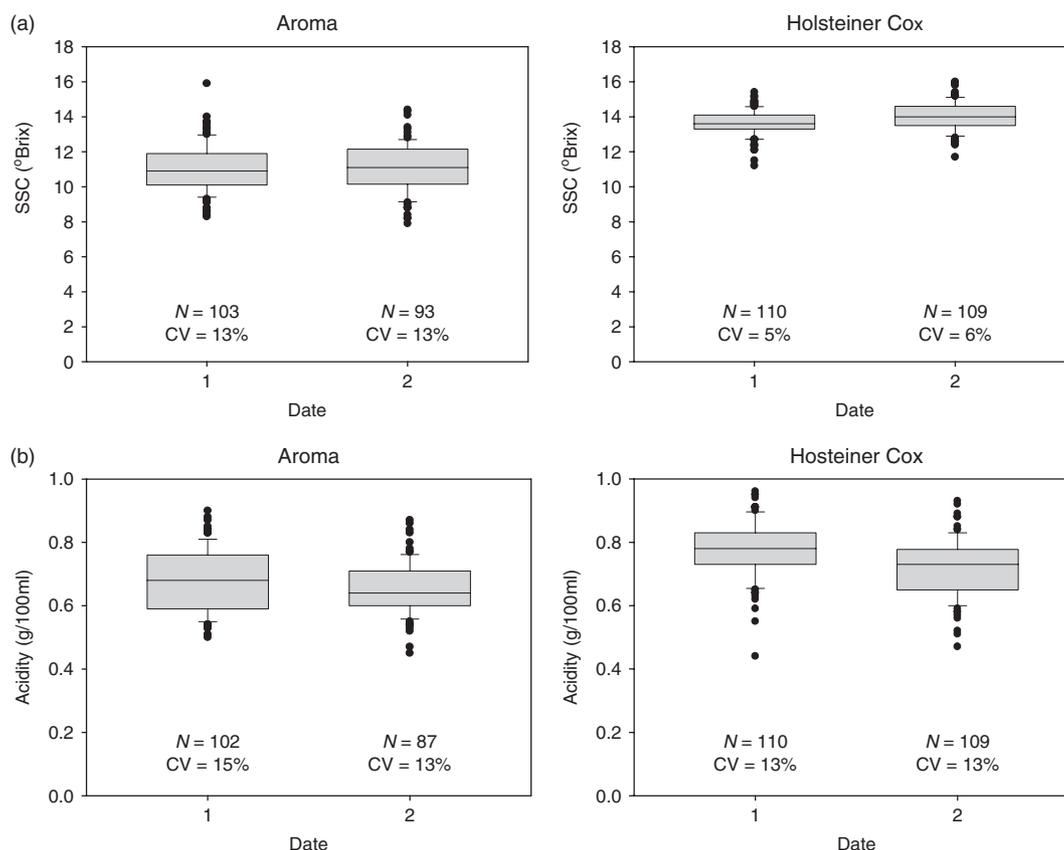
Figure 1. Box and whisker plots for (a) SSC (soluble solids content) and (b) acidity values for cultivars 'Aroma' and 'Holsteiner Cox' on two post-storage measurement dates (5 and 10 days). Extreme values, present for both variables and measurement dates are indicated by solid symbols on the plots. $N =$ number of samples; $CV = SD/mean$.

cultivars with potential for postharvest storage (unpublished data). The lower coefficient of variation ($CV = SD/mean$) of SSC for 'Holsteiner Cox' as compared to 'Aroma' , showed that 'Holsteiner Cox' samples had a notably narrower spread of SSC values (Fig. 1a). 'Aroma' and 'Holsteiner Cox' had almost similar CVs of acidity on both harvest dates.

Spectral signatures of the early and late season cultivar

The spectral signatures for the apple cultivars 'Aroma' and 'Holsteiner Cox' obtained in two different post-storage dates are illustrated in Fig. 2.

There were differences in the shapes of the spectral signature between cultivars and between measurement dates. Furthermore, the curves showed large variability in absorbance at a given wavelength. The visible region (below 700 nm) of the spectra appeared more irregular than the NIR region (above 700 nm) between Dates 1 and 2.

The 'Aroma' signature showed a noisy area in the blue region 400–500 nm. On the further green region 500–600 nm, there were differences on the turning points of the curve between Date 1 and Date 2 spectra. Different spectral regions have been related to chemical components such as chlorophyll at 650–695 nm[18] or carotenoids and anthocyanins at shorter wavelengths than 650 nm, sugars in 470–484 nm, 498–512 nm, 526–540 nm, 568–582 nm, 665–679 nm,[19] and sour taste (acidity) in the 640–700 nm region.[20] The low absorption values around the area between 700 and 900 nm probably do not contain important information for 'Aroma' and 'Holsteiner Cox' cultivars. In this

sense, spectral bands with almost zero light absorption have been reported to be influenced mainly by scattering properties of the tissue.[21] These spectral regions were not removed for the model calculations, however.

The peaked-shaped area shape above 900 nm is consistent with previous studies of SSC on apples. One should expect to find spectral curve peaks at around 800 nm related also to chlorophyll content,[22] 950 nm peaked areas may be related to water content and sugar–water peaks at 840 and 890 nm[23] and the overtones of the hydroxy (O—H) stretch/vibration of H_2O/carbohydrates may be explained at 930–1080 nm as well as variations in the absorption at 960 and 1060 nm, which are related to absorption of pure water and solutions of different sugar concentrations.[18]

Results of band selection

Figure 3 shows the counts of non-zero coefficients for each of the 731 wavelengths. The plot corresponds to the analysis of the original SSC data without MSC pre-treatment.

Figure 4 shows the resulting averages of the RMSEP values plotted after EN was applied on bands of training data according to the candidate list. As described previously in the feature selection section.

For SSC, the minimum RMSEP occurred at 600 features, but there was a very close RMSEP value also at 350. Because 350 bands was considerably smaller than 600, the first 350 top bands were selected for SSC. The same procedure was performed for acidity. In this case, the first 250 bands were chosen. The selected bands for SSC and acidity are shown in Fig. 5.
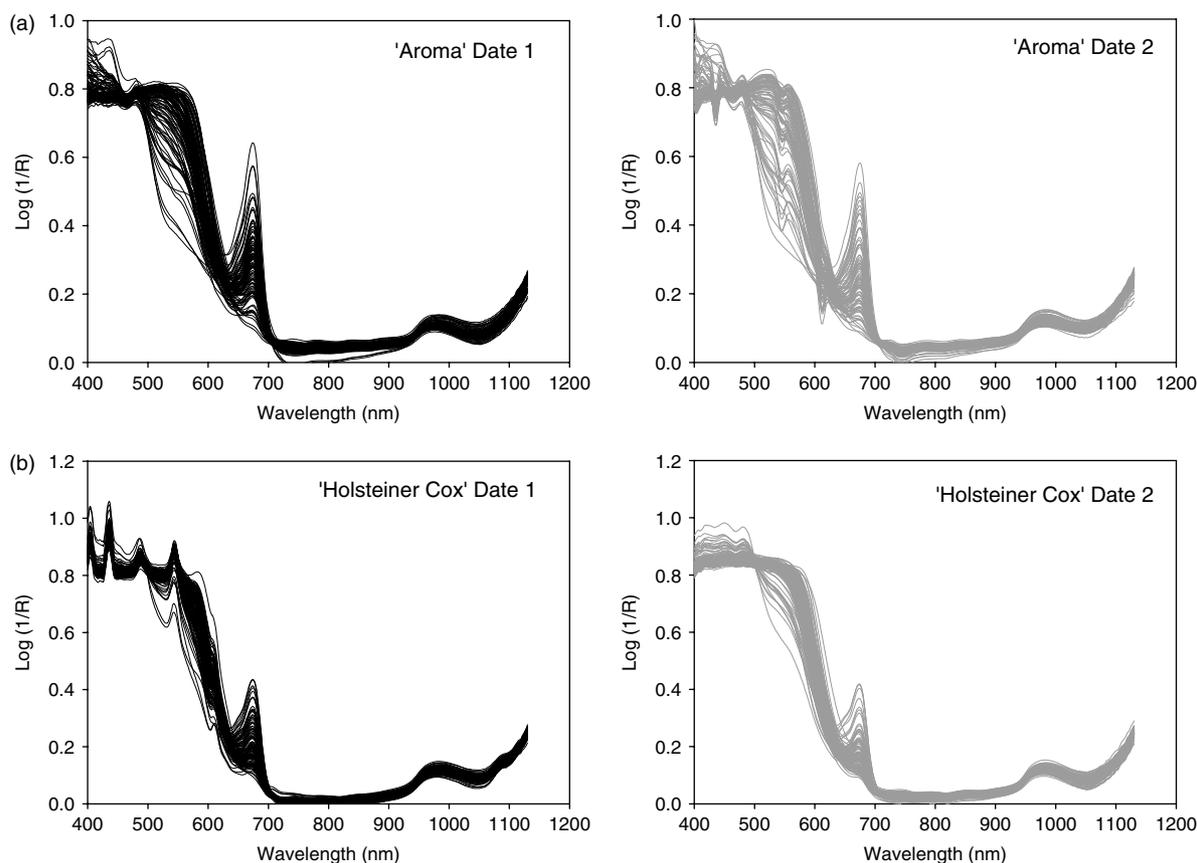
Figure 2. Raw spectral patterns recorded in the visible–NIR region 400–1100 nm (exposed and non-exposed sides of the fruit averaged and MSC pre-processed) and expressed as 'Absorbance' for (a) 'Aroma' and (b) 'Holsteiner Cox' in two measurement dates. Axes: X = wavelength (400–1100 nm), and Y = absorbance.
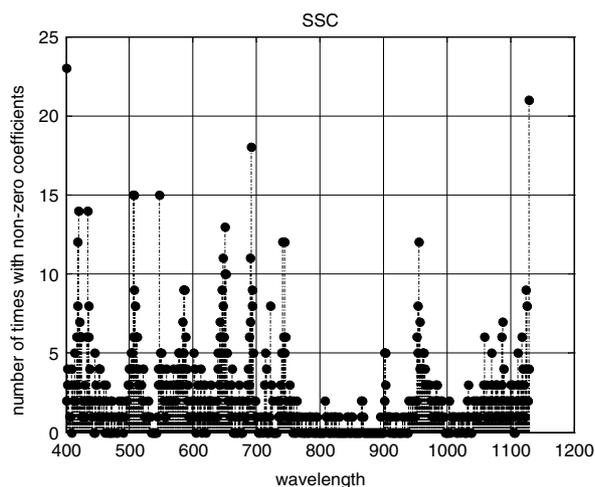
Figure 3. The frequency of having non-zero regression coefficients in 25 iterations of EN for the original SSC data set for 'Aroma' apples.
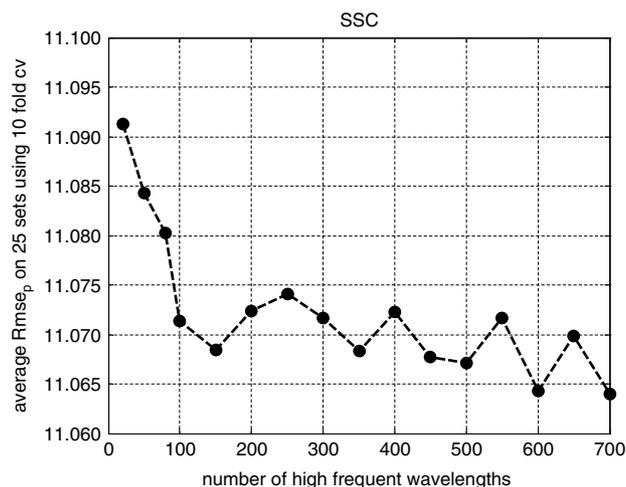
Figure 4. The RMSEP candidate number for SSC.

All the described steps were also performed with the MSC pre-processed data. The number of selected bands for SSC and acidity in this case were 450 and 250 respectively.

Results for the calibration and validation models

The resulting numbers of fruit samples on each of the previously explained sampling arrangements were: Sets A and C of 'Aroma'
had 147 and 49 (196 in total) samples for SSC and 141 and 48 (189 in total) samples for acidity. 'Holsteiner Cox' SSC training and test sets A and C had 165 and 54 (219 in total) respectively and the sets for acidity had 152 and 51 (203 in total) samples.

The smallest RMSEPs from each combination of the three arrangements and two data sets (original/MSC) are presented in Fig. 6. The selected features on Set A (smooth arrangement) using the EN regression and 26 wavelengths obtained the best
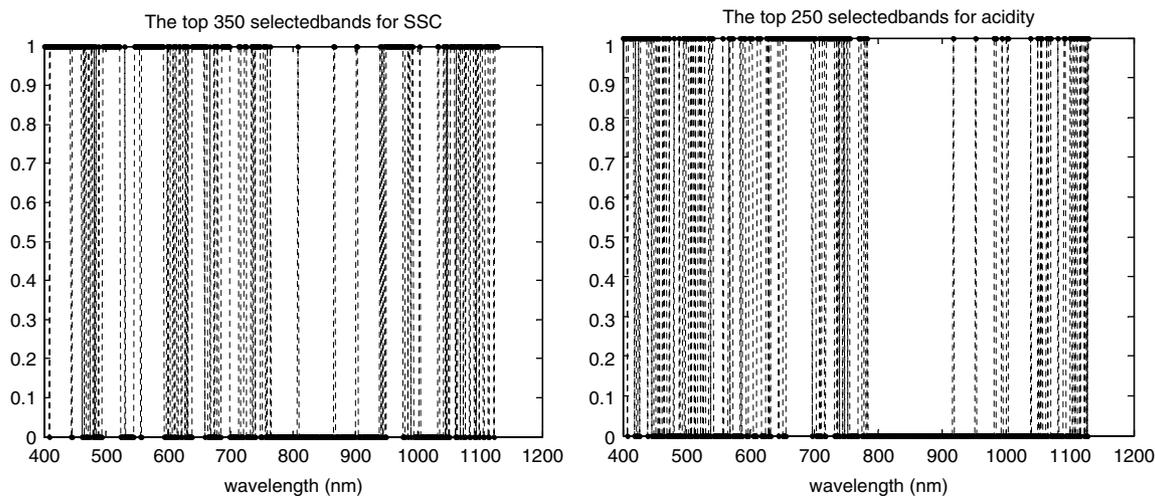
Figure 5. Stem plots for the selected bands for SSC and acidity for the original data set.
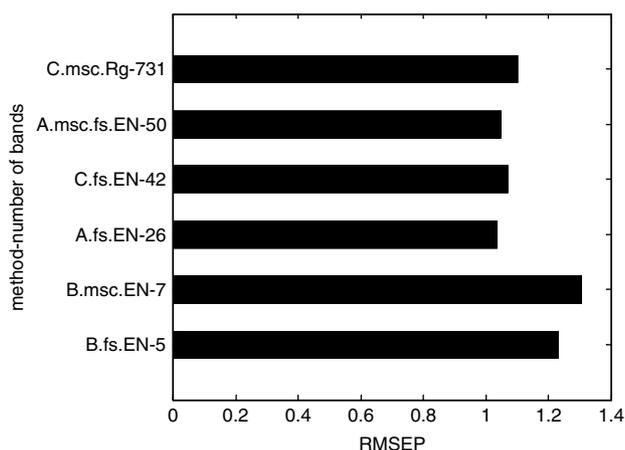
Figure 6. Overall comparison of the soluble solids content (SSC) prediction errors based on RMSEPs of the three training/test arrangements (A, B and C respectively) and the two original (without) and MSC pre-processed data. MSC, multiplicative scatter correction; Rg, ridge regression; FS, feature selection; EN, elastic net. The numbers on the ordinate indicate the number of wavelengths used in the models.

results for SSC and acidity prediction. Therefore complete results of sets B and C are not presented.

Figure 7 illustrates a comparison of the error of the prediction models obtained on sets B (Fig. 7a) and C (Fig. 7b) respectively. The figure further demonstrates the importance of the strategy used for forming the training and test sets. The minimum RMSEPs for set B were higher than the worst results obtained using set A using PLS. Set C also produced better models than Set B, but the best results were not as good as those for Set A.

Sets B and C had higher prediction errors. In particular the random sets (C) often caused over fitting during the modelling process, resulting often in poorer models. As an illustration, Fig. 8 shows the spread of RMSEP from the three sample arrangements used for building the prediction models of SSC of 'Aroma' apples. Summary statistics for the four sets and their average formed during smooth arrangement (Set A) are shown in Table 1.

Table 2 presents the average prediction statistics for the set 'A' of the 'Aroma' cultivar, which obtained the best results for both SSC and acidity models. In general, the calibration and prediction

correlation using PLS were inferior to RR and EN in all cases (Table 2 and Table 3). The error and bias remained low for set A.

In a similar manner, Table 3 shows the ridge and EN regression results for the selected bands of the original and MSC pre-processed data. For the SSC data, in all cases except ridge regression on the original data, the performance slightly improved using the reduced number of wavelengths. In the case of acidity, the effect was the same.

Soluble solids content

Table 2 shows that elastic net and ridge regression improved the prediction of SSC and also resulted in lower errors and bias as compared to PLS. Table 3 demonstrates that even though all the models performance in terms of R^2, errors and bias were not importantly improved after band selection, fewer bands on the visible and NIR region were suited for SSC prediction for set A. Fewer bands simplify the measurement systems and make them more cost effective. Our 'Aroma' SSC calibration models performed better ($R^2 = 0.44$; SEC $= 0.88°$Brix; range: 8.0–15.5$°$Brix; SD $= 1.39$) than previous studies done by Zude et al.,[21] which reported $R^2 = 0.04$ and higher SEC $= 1.82$ (for stored 'Golden Delicious' apples). They used higher number of samples in storage ($n = 250$; SD $=$ not reported) and spectral bands between 400 and 1000 nm. On the other hand, Dai et al.[24] obtained SSC prediction models with higher R^2 using smaller sample numbers ($N = 58$; $R^2 = 0.76$; SEC $= 0.22$; SEP $= 0.83$), similar band range 400–1100 nm for a data set with SSC values fairly normally distributed around the mean (range$_{cal} = 8.6–16.7$; range$_{val} = 8.6–15.5$ SD$_{cal} = 1.69$; SD$_{val} = 1.62$). The high difference between SEC and SEP in this latter study indicates that the training and test samples were not very similar. Another model from Hernández et al.[25] also had high prediction results ($R^2 = 0.98$; SEC $= 0.45°$Brix; SEP $= 1.69°$Brix) except bias $= 1.62$ was quite high. They used samples with higher variation than ours (CV$_{SSC} = 0.28$). It was not clear, however, how the samples were collected in these latter studies. Peirs et al.[8] calculated a SSC model using 244 apple samples for calibration and 244 samples validation from seven different apple cultivars where $R^2_{cal} = 0.91$ to 0.92, SEC $= 0.49$ to 0.76 but using spectral regions between 380 and 2000 nm. It is possible that better results for SSC might be obtained by extending the spectral data to other areas of the NIR spectra to the ones we studied.
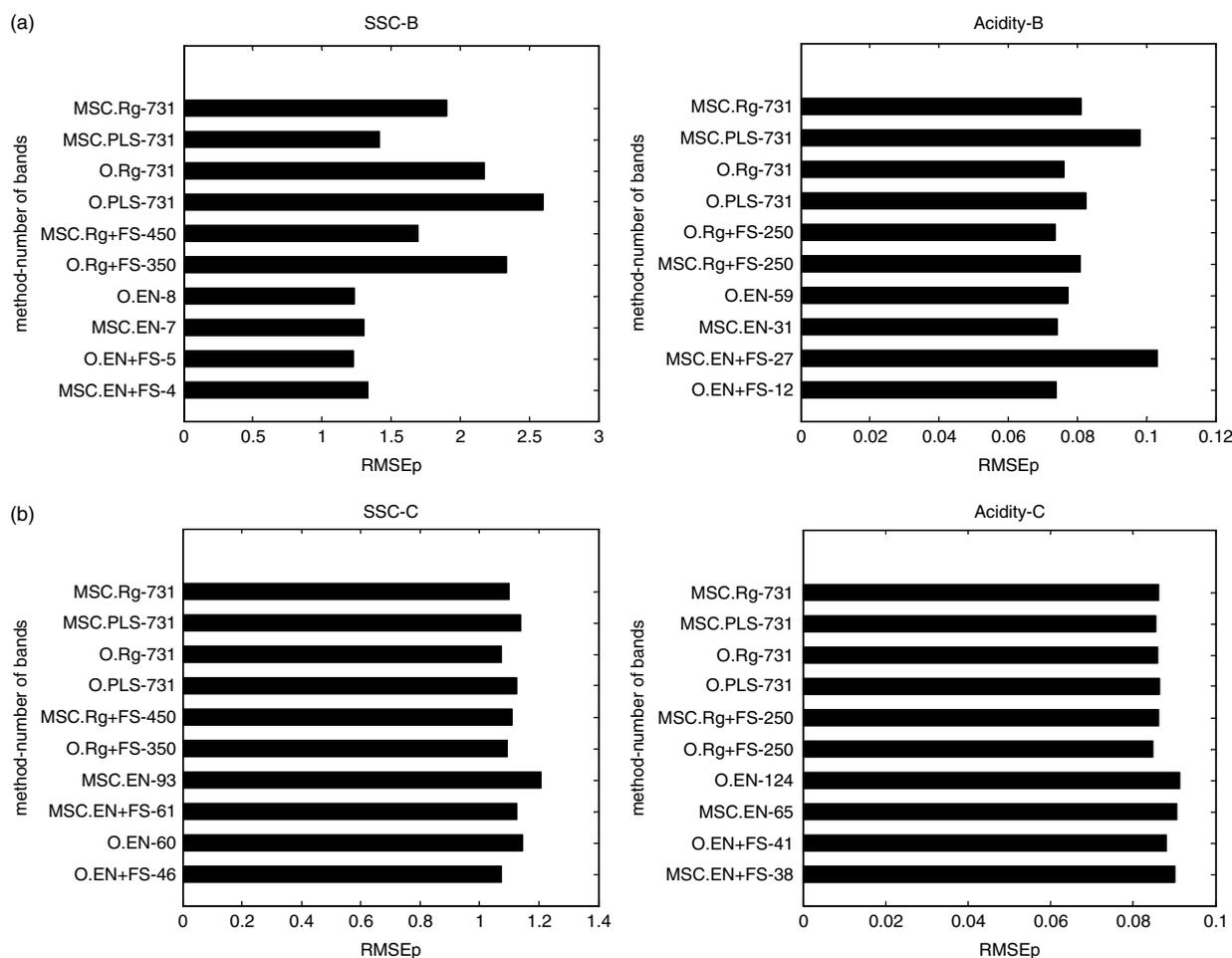
Figure 7. Comparison of the RMSEPs of the different regression methods used on the data division for modelling sub-samples (a) Set B and (b) set C for 'Aroma' cultivar. The numbers on the ordinate indicate the number of wavelengths used in the models and the letters indicate, in each case, the approach used to analyse the data sets. MSC, multiplicative scatter correction; O, original data without MSC; Rg, ridge regression; FS, feature selection; EN, elastic net.

Table 1. Statistics for the training/test sub-sample sets for the sample arrangement A

Characteristic	Statistic	Set 1	Set 2	Set 3	Set 4	Average
Soluble solids content (°Brix)	Number	147/49	147/49	147/49	147/49	147/49
	Range	7.9–14.4/8.3–15.9	7.9–15.9/8.2–14.3	7.9–15.9/8.2–14.4	8.2–15.9/7.9–14.1	8.0–15.5/8.1–14.7
	Mean	11.05/11.08	11.06/11.05	11.06/11.04	11.06/11.05	11.1/11.1
	SD	1.37/1.47	1.41/1.37	1.40/1.38	1.40/1.39	1.39/1.40
Acidity (g 100 mL^{-1})	Number	141/48	141/48	141/48	141/48	141/48
	Range	0.5–1.1/0.45–0.87	0.47–1/0.45–0.88	0.45–0.91/0.5–0.1	0.45–1/0.51–0.9	0.47–0.98/0.48–0.91
	Mean	0.67/0.67	0.67/0.67	0.67/0.67	0.67/0.67	0.67/0.67
	SD	0.09/0.01	0.09/0.1	0.09/0.1	0.1/0.09	0.09/0.1

Acidity

Our acidity models performed lower those reported by Peirs *et al.*[8] They used random samples from a combination of seven different apple cultivars to construct calibration ($N = 244$) and validation ($N = 244$) models in the region between 380 and 2000 nm. Their results were $R^2_{cal} = 0.88$, $R^2_{val} = 0.86$ (SEC = 1.64; SEP = 1.73). Abu-Khalaf and Bennedsen[26] reported also better results using in total 200 samples of two apple cultivars ('Golden Delicious' and 'Jonagold') to calculate calibration ($N = 130$; $r^2 = 0.84$; SEC = 0.07) and validation models ($N = 70$; SEP = 0.07)

on the spectral region 400–1100 nm. Other studies have reported improved prediction results for acidity on citrus fruit using the spectral region between 500–1100 nm ($r^2 = 0.65$; RMSEP = 0.15) and up to the 2500 nm acidity predictions have reached $r^2 = 0.86$; RMSEP = 0.17.[27]

'Holsteiner Cox' models had poor prediction, which was somehow expected because the sample variability from this cultivar was very low (CV of 5–6%). However, this also means that the spectra may be did not capture SSC or acidity levels accurately for this cultivar.

Table 2. Averaged calibration and prediction results (Set A) for 'Aroma' apple cultivar using all 731 spectral bands

Characteristic	Statistic	Raw data sets			MSC transformed data sets		
		PLS	RR	EN	PLS	RR	EN
Soluble solids content (°Brix)	N_{cal}	147	147	147	147	147	147
	N_{val}	49	49	49	49	49	49
	RMSECV	0.91	0.78	0.90	0.91	0.87	0.85
	RMSEP	1.09	1.04	1.04	1.11	1.08	1.05
	R^2_{cal}	0.55	0.68	0.56	0.60	0.60	0.61
	R^2_{val}	0.33	0.43	0.43	0.35	0.39	0.43
	SEC	0.93	0.79	0.92	0.92	0.88	0.86
	SEP	1.07	1.04	1.04	1.11	1.08	0.99
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	0.01	0.02	0.01	−0.03	0.01	0.02
	NNC	731	731	45	731	731	134
Acidity (g 100 mL^{-1})	N_{cal}	141	141	141	141	141	141
	N_{val}	48	48	48	48	48	48
	RMSECV	0.08	0.08	0.07	0.08	0.08	0.08
	RMSEP	0.08	0.08	0.08	0.09	0.08	0.08
	R^2_{cal}	0.29	0.30	0.36	0.29	0.27	0.33
	R^2_{val}	0.22	0.22	0.22	0.15	0.18	0.18
	SEC	0.08	0.08	0.08	0.08	0.08	0.08
	SEP	0.09	0.09	0.08	0.09	0.09	0.09
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
	NNC	731	731	81	731	731	26

MSC, multiplicative scatter correction.
PLS, partial least squares method.
RR, ridge regression.
EN, Elastic Net.
N_{cal}, number of samples in the calibration set.
N_{val}, number of samples in the prediction set.
RMSECV, root mean square error of cross validation.
RMSEP, root mean square error of prediction.
R^2_{cal}, coefficient of determination (calibration).
R^2_{val}, coefficient of determination (validation).
SEC, standard error of calibration.
SEP, standard error of prediction.
$Bias_{cal}$, bias calibration.
$Bias_{val}$, bias validation.
NNC, number of non-zero coefficients.

The differences on model robustness observed between the three approaches used for forming training and test sets, indicated that training and test sample arrangement do affect the overall model performance, especially when the number of samples are limited and smaller than the number of wavelengths. The tendency of producing stable prediction after applying the smooth fractionator to form training and test sets, stressed the importance of maintaining the original sample variability throughout the entire model construction process in order to achieve model robustness and high precision (higher coefficients of determination, low errors between calibration and validation sets and low bias). This conclusion is supported by the differences observed between our training and test sets performance, which suggests that the variability of the training set were, by chance, excluded from the test set during the formation of training and test sets. Consequently, using a different sampling period ($p = 2$) to form training and test sets, so that the proportion becomes 50–50 instead of the commonly used 75–25, might be a better approach to use in order to preserve as much as possible the original variability of the whole data set, when working with smaller samples. The fractionator technique used to sample fruit from the trees has already been shown to be an effective way to obtain small samples (<100) representing the variability of fruit size and internal quality at the orchard scale, as shown by Wulfsohn et al.[28] and Martinez V et al. (unpublished). These previous studies and the findings of this study suggest that is feasible to develop robust visible–NIR prediction models using relatively small samples. The type of cross validation used might have some additive effect on the model performance as well, because it is a method also based on repetitive selection of samples from the calibration set.[29] However, given the low results for the model errors and bias, we consider the robustness of our models to be adequate for this type of data set.

In practice, the results of this study imply that using local fruit samples for developing spectral calibration and prediction sets is feasible, as long as the sample variability is taken into account on the formation of training and prediction sets. The late season samples, however, need to be modelled differently. Using representative fruit samples with higher internal quality variability (e.g. CV > 15%) might be a better approach to use, since a much
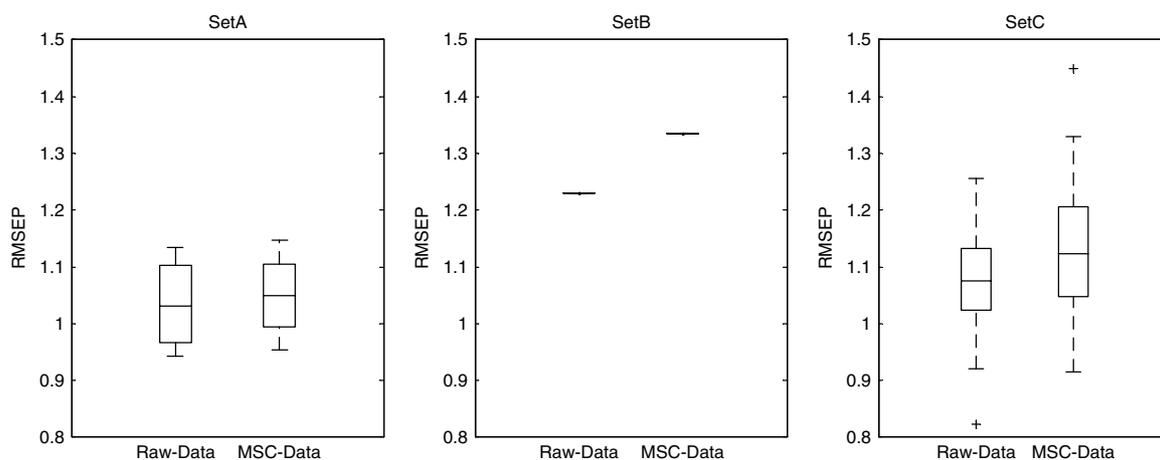
Figure 8. Box plots for the root mean square of prediction (RMSEP) for the raw and MSC treated data of the sample arrangements A, B and C.

Table 3. Averaged calibration and prediction results (set A) for Aroma apple cultivar using selected bands

Characteristic	Statistic	Raw data sets			MSC transformed data sets		
		PLS	RR	EN	PLS	RR	EN
Soluble solids content (°Brix)	N_{cal}	147	147	147	147	147	147
	N_{val}	49	49	49	49	49	49
	RMSECV	0.87	0.77	0.88	0.92	0.88	0.87
	RMSEP	1.14	1.07	1.03	1.13	1.07	1.05
	R^2_{cal}	0.60	0.69	0.60	0.55	0.60	0.61
	R^2_{val}	0.33	0.40	0.44	0.33	0.40	0.43
	SEC	0.88	0.78	0.88	0.93	0.89	0.87
	SEP	1.14	1.07	1.03	1.13	1.08	1.05
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	0.04	0.03	0.03	0.00	0.02	0.02
	NNC	350	350	26	450	450	50
Acidity (g 100 mL^{-1})	N_{cal}	141	141	141	141	141	141
	N_{val}	48	48	48	48	48	48
	RMSECV	0.08	0.08	0.08	0.08	0.08	0.08
	RMSEP	0.08	0.08	0.09	0.08	0.08	0.08
	R^2_{cal}	0.32	0.31	0.27	0.32	0.31	0.30
	R^2_{val}	0.22	0.24	0.19	0.19	0.20	0.18
	SEC	0.08	0.08	0.08	0.08	0.08	0.08
	SEP	0.09	0.08	0.09	0.08	0.08	0.09
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
	NNC	250	250	37	250	250	19
Abbreviations as in the footnote to Table 2.							

higher number of samples of the fruit populations with the same variability as the samples we presented will not necessarily improve the prediction results greatly. Such high variability samples may be obtained from commercial orchards where the internal variability of fruits is expected to be higher. We are testing this hypothesis in different batches in our further research.

CONCLUSION

Three sub-sampling techniques for the formation of training and test sets for spectral prediction of SSC and acidity of an early apple cultivar were compared. The smooth fractionator approach was clearly superior to random sampling and to by-date separation,

resulting in models with consistently low errors and low bias, mainly because the method provides a fair representation of the response values in both the training and test sets. In addition, three different methods to reduce model complexity, PLS, RR and EN were compared. Using elastic net and fewer bands were the best approaches to reduce model complexity ($R^2 = 0.44$; SEP = 1.03°Brix; bias = 0.03; range: 8.1–14.7°Brix). Therefore our results confirmed that the fractionator sampling provide data sets suitable for SSC prediction with visible–NIR spectroscopy. To our knowledge, this is the first proposal of a modelling protocol for a sub-sample of training and test sets, which takes into account the variability of the original sample in the context of predicting fruit quality by using a non-destructive method.

ACKNOWLEDGEMENTS

REFERENCES

1 Cozzolino D, Cynkar WU, Shah N and Smith P, Multivariate data analysis applied to spectroscopy: potential application to juice and fruit quality. *Food Res Int* **44**:1888–1896 (2011).
2 McCaffrey D, Harvesting the sun. A profile of world horticulture. *Script Hortuc* **14** (2012).
3 Nicolai B, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron K, *et al.*, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biol Technol* **13**:99–118 (2007).
4 Mehinagic E, Royer G, Bertrand D, Symoneaux R, Laurens F and Jourjon, Relationship between sensory analysis, penetrometry and visible–NIR spectroscopy of apples belonging to different cultivars. *Food Qual Prefer* **14**:473–484 (2003).
5 Mendoza F, Lu R and Cen H, Comparison and fusion of four non-destructive sensors for predicting firmness and soluble solids content. *Postharvest Biol Technol* **73**: 89–98 (2012).
6 Bobelyn E, Serban A, Nicu M, Lammertyn J, Nicolai B and Wouter S, Postharvest quality of apple predicted by NIR-spectroscopy: study of the effect of biological variability on spectra and model performance. *Postharvest Biol Technol* **55**:122–143 (2010).
7 McGlone VA and Kawano S, Firmness, dry-matter and soluble solids assessment of postharvest kiwifruit by NIR spectroscopy. *Postharvest Biol Technol* **13**:131–141 (1998).
8 Peirs A, Lammertyn J, Ooms K and Nicolai BM, Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR spectroscopy. *Postharvest Biol Technol* **21**:189–199 (2000).
9 Wulfsohn D, Maletti G and Toldam-Andersen TB, Unbiased estimator of the total number of flowers on a tree. *Acta Hortic* **707**:245–251 (2006).
10 Næs T, Isaksson T, Fearn T and Davies T, *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester, pp. 114–115 (2002).
11 Gundersen HJG, The smooth fractionator. *J Microsc* **207**:191–210 (2002)
12 Wulfsohn D, Sampling techniques for plants and soil. *Landbauforschung Volkenrode* **340**:3–30 (2010).
13 Hastie T, Tibshirani R and Friedman J, *The Elements of Statistical Learning*, 2nd edition. Springer, Canada, pp. 61–68 (2009).
14 Zou H and Hastie T, Regularization and variable selection via Elastic Net. *J R Statistics Soc* **67**:301–320 (2005).
15 Hoyle D C and Rattray M, Statistical mechanics of learning multiple orthogonal signals: Asympotic theory and fluctuation effects. *Phys Rev E (Statistical, Nonlinear and Soft Matter Physics)*, **75**:016101 (2007).
16 Sharifzadeh S, Clemmensen HL, Borggaard C, Støier S and Ersbøll KB, Supervised Feature Selection for Linear and Non-Linear Regression of L*a*b* Color from Multispectral Images of Meat, Accepted at Eng App Artif Intel, Manuscript Number: EAAI-12-1827 (2013).
17 Korsgaard M and Toldam-Andersen TB, *The Apple Key*. Available: http://www.nordgen.org/nak/index.php?view=show&id=5003 [7 August 2012].
18 Qing Z, Bi J and Zude M, Wavelength selection for predicting physicochemical properties of apple fruit based on near-infrared spectroscopy. *J Food Qual* **30**:511–526 (2007).
19 Xiaobo Z, Yanxiao L and Jiewen Z, Using genetic algorithm interval partial least squares selection of optimal near infrared wavelength regions for determination of the soluble solids content of 'Fuji' apple. *J Near Infrared Spectrosc* **15**:153–159 (2007).
20 Rizzolo A, Vanolli M, Spinelli L and Torricelli A, Sensory characteristics, quality and optical properties measured by time reflectance spectroscopy in stored apples. *Postharvest Biol Technol* **58**:1–12 (2010).
21 Zude M, Herold B, Roger JM, Bellon-Maurel V and Landahl S, Non-destructive tests on the prediction of apple fruit flesh firmness and soluble solids content on tree and in shelf life. *J Food Eng* **77**:254–260 (2006).
22 Sun D, *Contemporary Food Engineering*. CRC Press, New York, chapter 3 (2009).
23 Chauchard F, Roger JM and Bellon-Maurel V, Correction of the temperature effect on near infrared calibration–application to soluble solid content prediction. *J Near Infrared Spectrosc* **12**:199–205 (2004).
24 Dai F, Hong T, Zhang K and Chen H, Comparison of modelling methods in rapid estimation of sugar content of apple based on near infrared spectrum. An ASABE meeting presentation, paper number: 1008620. ASABE, St. Joseph, Mich. (2010).
25 Hernández N, Lurol S, Roger J and Bellon-Maurel V, Robustness of models based on NIR spectra for sugar content prediction in apples. *J Near Infrared Spectrosc* **11**:97–107 (2003).
26 Abu-Khalaf N and Benenedsen B, Near infrared (NIR) technology and multivariate data analysis for sensing taste attributes of apples. *Int Agrophys* **18**:203–211 (2004).
27 Magwaza LS, Opara UL, Niewoudt H, Cronje PJR, Saeys W and Nicolaï B, NIR spectroscopy applications for internal and external quality analysis of citrus fruit – A review. *Food Bioproc Technol* **5**:425–444 (2012).
28 Wulfsohn D, Aravena F, Potin C, Zamora I and Garcia-Fiñana M, Multilevel systematic sampling to estimate total fruit number for yield forecasts. *Precis Agric 13*, **2**:256–275 (2012).
29 Gemperline P, *Practical Guide to Chemometrics*, 2nd edition. Taylor and Francis, Boca Raton, pp. 115–116 (2006).