

Assessment of internal quality attributes of mandarin fruit. 2. NIR calibration model robustness

J. A. Guthrie^{A,B,C}, D. J. Reid^A, and K. B. Walsh^B

^ADelivery, Queensland Department of Primary Industries and Fisheries, PO Box 6014, CQ Mail Centre, Rockhampton, Qld 4702, Australia.

^BPlant Sciences Group, Central Queensland University, Rockhampton, Qld 4702, Australia.

^CCorresponding author. Email: john.guthrie@dpi.qld.gov.au

Abstract. The robustness of multivariate calibration models, based on near infrared spectroscopy, for the assessment of total soluble solids (TSS) and dry matter (DM) of intact mandarin fruit (*Citrus reticulata* cv. Imperial) was assessed. TSS calibration model performance was validated in terms of prediction of populations of fruit not in the original population (different harvest days from a single tree, different harvest localities, different harvest seasons). Of these, calibration performance was most affected by validation across seasons (signal to noise statistic on root mean squared error of prediction of 3.8, compared with 20 and 13 for locality and harvest day, respectively). Procedures for sample selection from the validation population for addition to the calibration population ('model updating') were considered for both TSS and DM models. Random selection from the validation group worked as well as more sophisticated selection procedures, with approximately 20 samples required. Models that were developed using samples at a range of temperatures were robust in validation for TSS and DM.

Additional keyword: non-invasive.

Introduction

In a companion manuscript (Guthrie *et al.* 2005) the reference sampling procedure and data pre-processing techniques were optimised for the development of partial least squares (PLS) calibration models on intact mandarin fruit for total soluble solids (TSS) and dry matter (DM), using short wavelength (720–950 nm) near infrared (NIR) spectra acquired in an interactance mode (Greensill and Walsh 2000). Chemometric descriptive terms were also defined and will be used in the current manuscript.

However, the application of near infrared spectroscopy (NIRS) to a given fruit commodity requires an assessment of the robustness of the calibration model across populations of fruit grown under differing conditions. Different growing conditions may result in differences in physical (e.g. trichome density, intercellular space content) and chemical (e.g. water content) properties of a fruit, resulting in altered fruit optical properties and band assignments. Unfortunately, most reports on the application of NIRS to fruit sorting describe the use of a single harvest population, divided into a calibration set and a validation set. For example, McGlone *et al.* (2003) worked with 20 populations of mandarins (from 3 orchards over 7 weeks), with a calibration set assembled from 75% of each set and used to predict a population consisting of the remaining samples from each of the sets. This procedure

allowed an estimate of prediction error; however, it did not involve testing the calibration against fruit from independent populations (e.g. different harvest dates or localities). We are aware of only 3 relevant reports involving the use of separate harvest populations of fruit for calibration and validation. These studies involved mandarin and peach fruit.

Two of these studies on TSS model robustness for fresh fruit have involved mandarin. Ou *et al.* (1997) reported the use of a calibration developed in 1 fruit-growing region to predict TSS of Ponkan mandarin fruit from that region and from 2 other regions, with coefficient of determination on validation data set (R_v^2) of 0.72, 0.44, and 0.30, and root mean squared error of prediction (RMSEP) of 0.68, 1.16, and 1.28% TSS, respectively. A calibration based on data combined across regions performed better, with R_v^2 of 0.76 and RMSEP of 0.92% TSS. Miyamoto and Kitano (1995) reported the use of a calibration developed in 1 year to predict TSS content of intact Satsuma mandarins in the subsequent 2 seasons. Prediction statistics were similar to those for calibrations developed within a given season (RMSEP of <0.6% TSS and bias of $\leq 0.4\%$ TSS).

The third published study on TSS model robustness for fresh fruit involved peach. Miyamoto and Kitano (1995) also reported calibration validation across 3 seasons for peach. Prediction statistics for a calibration developed across data

from all years (RMSEP of 0.60% TSS, *bias* of <0.1% TSS) were better than for a calibration developed in any 1 year (RMSEP of 0.64% TSS, *bias* up to 0.34% TSS). Also using peaches, Peiris *et al.* (1998) reported that a calibration developed on a population drawn from 3 seasons predicted better on a combined season validation set (RMSEP of 0.9–1.3% TSS and *bias* of 0.2–0.4% TSS) than that developed from populations drawn from a single season (RMSEP of 0.9–1.4% TSS and *bias* of 0.2–2.1% TSS).

The comparisons of model validation among independent populations are usually difficult because population attributes (e.g. SD) vary. The standard deviation ratio (SDR), expressed as the ratio of SD to root mean squared error of cross validation (RMSECV) (for calibration datasets) or RMSEP (for validation datasets), or RPD (ratio of the SD to RMSECV corrected for *bias* [RMSECV(C)] or RMSEP corrected for *bias* [RMSEP(C)] of the data) (Williams and Sobering 1993), is sometimes presented as a gauge of the utility of the technique. Other indices have also been used. Ou *et al.* (1997) reported a form of the coefficient of variation (CV) statistic ($CV = \text{RMSEP}/\text{mean of the prediction set}$), and Miyamoto and Kitano (1995) reported an evaluation index ($EI = 2 * \text{RMSEP}/\text{range} * 100$), in an attempt to compare model performance across populations. Another approach, suggested by Wortel *et al.* (2001), is based on the Taguchi concept of process control, in which the variation of RMSEP among validation populations of a given condition (e.g. populations drawn from different harvest regions) is quantified in a signal to noise (S/N) statistic ($S/N = 20 * \log_{10} [\text{mean RMSEP}/\text{SD}_{\text{RMSEP}}]$, where mean RMSEP is the average of the RMSEP across all validation populations, and SD_{RMSEP} is the SD of all the RMSEP values).

In other industries (e.g. cereal, oilseed) NIRS-based models are extended by inclusion of samples from the validation population (e.g. from a new variety of wheat or a new season of oilseed production). The decision on when to add new samples to the calibration set is generally based on an assessment of the dissimilarity of the calibration and validation sets based on principal component analysis (PCA)/PLS scores. The Mahalanobis distance, D (Mahalanobis 1936), is such a measure. The chemometric software package WINISI (ver. 1.04a) uses mean-centred score data in the calculation of D . Further, D is normalised to f to create the global H (GH) statistic, as follows:

$$GH = \frac{D^2}{f}$$

where f is the number of PCA/PLS factors in the model.

Shenk and Westerhaus (1991) advocate the use of the GH value and a 'nearest neighbour' Mahalanobis distance (NH, Mahalanobis distance from any given sample to its nearest neighbour in principal component space) for the selection

of outliers and for sample addition. However, the choice of how many and which samples from the validation set should be added to the calibration set is vexatious. Typically, high leverage samples, which are not outliers, will be chosen, with the number required defined through trial and error (e.g. Wang *et al.* 1991).

Calibration model performance is affected by sample temperature primarily through the strong effect of temperature on H bonding and thus on the absorption bands related to OH (Golic *et al.* 2003). Therefore, model robustness should also be considered with respect to this variable. We hypothesise that calibration models for DM would be more sensitive to temperature than models for TSS, as DM models may reflect water content.

Kawano *et al.* (1995) noted that as sample fruit (peach) temperature increased, so did absorption at 841 and 966 nm (water bands), resulting in a *bias* in the prediction of TSS. Miyamoto and Kitano (1995) noted that when a calibration developed from spectra collected from mandarin fruit at 20°C was used to predict the same fruit at 6, 15, and 25°C, RMSEP [presumably RMSEP(C)] was constant but *bias* increased linearly with temperature. These researchers developed MLR models using 3 wavelength regions, noting 900–910 nm to be directly relevant to sugar, 740–755 or 840–855 nm to compensate for the optical path-length of the fruit, and 794 or 835 nm to compensate for the influence of fruit temperature. Both reports concluded that if the calibration model was developed with sample temperatures covering the range of future sample temperatures, then prediction accuracy was high. Sanchez *et al.* (2003) also noted that the influence of spectrometer and fruit (apple) temperature was mainly on *bias*, not RMSEP(C). However, the effect of spectrometer temperature on *bias* was more than twice that of fruit temperature.

The 'repeatability' file option of WINISI (ver. 1.04a) software may represent an alternative procedure for developing robustness in the model with respect to sample temperature. This procedure was developed for the calibration transfer between instruments, and depends on the collection of spectra of a few samples on different instruments. However, rather than add spectra directly to the calibration, the repeatability file adds 'difference' spectra (for each sample, scanned under different temperatures), with corresponding reference values of zero (Westerhaus 1991). The calibration algorithm is modified to give extra weight to these spectra.

In the current study, we report on the robustness of NIRS models for the evaluation of attributes related to eating quality (% TSS, % DM) of intact mandarin, and on procedures to select samples for addition to the calibration set. Calibration robustness across harvest time, location, and seasons for prediction of TSS, using the assessment methodology suggested by Wortel *et al.* (2001), is considered. Calibration

robustness for prediction of TSS and DM is also considered with respect to sample temperature. Calibration performance across instruments (e.g. as reported for Imperial mandarins by Greensill and Walsh 2002) will not be considered here.

Materials and methods

Plant material, reference analyses, and spectroscopy

Mandarin fruit (Imperial variety) were obtained following commercial harvest from orchards in Munduberra (25.6°S, 151.6°E), Bundaberg (24.9°S, 152.3°E), and Dululu (23.8°S, 150.3°E), Queensland.

Populations used in this study are the same as those used in the companion study for the 2001 season, with populations alphabetically named in chronological order as described in Guthrie *et al.* (2005). Additional populations from the 1999 and 2000 seasons were used in Fig. 1. Model robustness across harvest day, location, and season was evaluated for TSS using populations gathered from a single tree over 2 weeks, from 3 different locations, and from 4 seasons (from different locations) (see Table 1). For TSS, the calibration population was a combined J and K, with validation populations of M, E, G, and L. For DM, the calibration population was T, with validation populations of V, R, and S. Additionally, a model developed on a combined population made up of 2 populations per year from years 1999 and 2000 ($n = 307$, mean 9.9 and SD of 1.44% TSS) was used to predict on a separate population from year 2000 (mean 14.2 and SD 1.05% TSS).

Total soluble solids content of extracted juice and DM of fruit halves were determined as described in Guthrie *et al.* (2005). The procedures used to acquire spectra were also described in Guthrie *et al.* (2005). Briefly, spectra were collected over the wavelength range 720–950 nm using a NIR-enhanced Zeiss MMS1 spectrometer and a 100-W tungsten halogen light in the intertance optical configuration reported by Greensill and Walsh (2000) (0° angle between illumination and detected light rays, with detection probe viewing a shadow cast by the probe onto the fruit).

Chemometrics

The software package WINISI (ver. 1.04a) was used for chemometric analysis. Calibrations were developed using both step-wise multiple linear regression (MLR) and modified partial least squares regression (MPLS). The data pre-treatment options of first derivative, standard-normal variance, and detrend scatter correction, as recommended in the companion study (Guthrie *et al.* 2005), were adopted throughout the current study. The repeatability file option in WINISI was also considered as a method to improve prediction statistics across the different sample temperatures.

The criteria of Wortel *et al.* (2001) were applied to evaluate model robustness. This approach involved calculation of an average RMSEP and the S/N statistic for the performance of a given model across a range of validation populations.

A common approach for the improvement of calibration performance on a new validation population involves the addition of samples from the validation population to the calibration population. In this study, we extend the treatments reported by Guthrie and Walsh (2001). Each validation population was initially assessed for outliers as samples with a GH >3.0 using its own scores and loadings. These outliers were removed and the resulting data divided randomly into 2 groups, one group (two-thirds) retained as the validation population and the other group used for selection of samples for addition to the calibration population.

The following 3 approaches were used in the selection of samples from the validation population for addition to the calibration population.

- (1) Random: done twice using 2 different seeds to the random number generator.
- (2) Selected on GH value: selecting samples with either (a) minimal GH (i.e. spectrally similar to the 'mean' of the calibration population), (b) maximal GH (i.e. spectrally dissimilar to the 'mean' of the calibration population), or (c) spaced evenly on GH ranking (i.e. representative of the 'spread' of the calibration population).
- (3) Selection on the basis of NH using 2 methods: (a) NH cut-off (in which only samples with a NH value greater than the 'cut-off' value are chosen; this procedure is available as a WINISI software option, under 'Make and Use Scores', 'Select Samples From a Spectra File'), and (b) NH end (a manual implementation of (a), in which all samples were ranked manually in ascending order of magnitude for NH, with high NH value samples chosen).

Thus, in total, 6 methods for sample selection were trialled. In these exercises, the GH and NH values were calculated for validation population members using the scores and loadings of the calibration population. All validation populations were independent of the calibration populations.

Different population updating techniques were compared, as were different numbers of samples for model updating. This was trialled on different calibration and validation populations for the attributes of both TSS and DM (see Figs 1–5).

For 1 population of mandarin fruit (population T), spectra were collected of fruit at room temperature (22°C) and then the fruit equilibrated to 10°C and 30°C and rescanned. These fruit were then assessed (separate halves) for both TSS and DM. Calibration models were developed on a population of 70 samples (mean 9.6% and SD 1.51% for TSS, and mean 14.7% and SD 1.66% for DM), from spectra collected of these fruit at 10, 22, and 30°C. The prediction populations were based on a separate population of 34 samples (mean 9.8%, SD 1.64% for TSS, and mean 14.7%, SD 2.03% for DM), again with spectra collected of these fruit at 10, 22, and 30°C. The calibration models for TSS involved 5 terms, whereas those for DM involved 6 or 7 terms. Separate calibration models were developed on spectra of fruit at 10, 22, 30, 10, and 22°C, and all 3 temperatures combined (i.e. 5 models). The WINISI repeatability file option was also used, with all samples or the 4 samples with lowest GH values from the 22°C validation population. The significance ($P < 0.05$) of differences in RMSEP and bias were tested as described by Fearn (1996), using an automated spreadsheet (Guthrie *et al.* 2005).

Results

Calibration model robustness

A given TSS calibration model was used to predict populations over harvest day, location, and season (Table 1). The model used to predict populations over harvest day and location was based on the combination of 2 populations (J and K), and the model for predicting populations over seasons (years) was based on populations from 1999. Model predictions were more variable across seasons than across harvest days or location (in terms of both RMSEP and bias). This prediction variability was indexed as an average RMSEP and a S/N on RMSEP following the procedure of Wortel *et al.* (2001). The S/N ratio on the RMSEP of the

Table 1. Calibration (Cal) and validation (Val) statistics for MPLS and MLR calibration models for mandarin TSS, with validation across several populations varying in day of harvest, harvest location, and season of harvest
Variation in prediction performance is reported in terms of the Taguchi S/N value, RMSEP, and average RMSEP. In the MPLS models the number of factors varied from 7 to 11. Population identifiers (letters in parentheses) refer to table 6 in the companion manuscript (Guthrie *et al.* 2005). R_c^2 , coefficient of determination on calibration data set

Fruit population	SD (% TSS)	R_c^2	MPLS RMSECV/RMSEP (% TSS)	<i>bias</i> (% TSS)	R_c^2	MLR RMSEP(C) (% TSS)	<i>bias</i> (% TSS)
<i>Harvest days</i>							
Cal (J–K) (Days 1 & 3)	0.92	0.90	0.33		0.86	0.36	
Val							
Day 5 (L)	1.04	0.86	0.41	0.06	0.79	0.54	–0.07
Day 7 (M)	0.74	0.56	0.53	–0.13	0.64	0.48	0.03
Day 9 (N)	0.68	0.56	0.48	0.11	0.49	0.56	0.02
Day 10 (P)	0.84	0.84	0.51	0.39	0.80	0.95	0.85
Day 13 (Q)	0.67	0.68	0.51	0.33	0.63	0.87	0.75
S/N RMSEP			19.8			10.1	
Av. RMSEP			0.49			0.68	
<i>Location</i>							
Cal (J–K)	0.92	0.90	0.33		0.86	0.36	
Val							
A (E)	0.99	0.75	0.59	0.31	0.6	0.79	0.39
B (F)	0.49	0.35	0.81	0.66	0.14	1.24	1.09
C (G)	0.60	0.53	0.95	0.85	0.30	1.31	1.15
S/N RMSEP			12.7			12.0	
Av. RMSEP			0.78			1.12	
<i>Seasons</i>							
Cal (1999)	0.92	0.93	0.27		0.87	0.34	
Val							
Year 1 (1999)	1.05	0.35	4.94	3.21	0.32	3.70	1.51
Year 2 (2000)	1.05	0.83	2.10	2.05	0.78	2.13	2.07
Year 3 (2001)	0.78	0.03	6.76	–3.58	0.02	4.74	–0.35
Year 4 (2004)	1.32	0.74	0.77	–0.31	0.21	2.74	1.95
S/N RMSEP			3.82			1.50	
Av. RMSEP			3.64			3.33	

MPLS model predictions was 20 over harvest days, 13 over location, and 4 over seasons (Table 1). Modified partial least squares models were more robust than MLR models (MLR models had lower S/N ratios, being 10, 12, and 2 for harvest days, locations, and seasons, respectively).

Model performance in prediction of TSS of an independent population was improved by inclusion of samples from the independent population, regardless of the method used to select the samples for inclusion (Figs 1 and 2). This was demonstrated for a calibration developed on populations from 1999 and 2000 (Fig. 1), and from 2001 (Fig. 2). The 4 methods of sample selection used (random, every 'ith' sample based on ranking by GH, maximum GH, and maximum NH) all behaved similarly (Fig. 1). Model performance improved from 1.1% TSS to 0.45% TSS for RMSEP and from 1.1% TSS to 0.15% TSS for *bias* with the inclusion of only 10 samples (Fig. 1). In the second exercise, where 6 methods of selection were used (as above

plus minimum GH and WINISI sample addition facility), all methods again behaved similarly in terms of *bias* (Fig. 2). In terms of RMSEP, all methods behaved similarly with addition of up to 5 samples, but there was some divergence among methods with addition of 10–20 samples. The RMSEP values increased with the addition of 10 samples for the 'random' and 'greatest GH' (approaches 1 and 2, respectively) and for the addition of 10 and 20 samples for the WINISI 'sample addition facility' (approach 3). However, a repeat of the random selection approach gave divergent results for the addition of 10 samples. Of course, with the addition of all samples from the one-third validation population, all results will converge (except for where there is a slight difference in the size of the population, e.g. 30 drawn from a population of 31 or 34).

Model performance (from calibration on Population T) in prediction of DM of an independent population (Population V) was also improved (in terms of both RMSEP

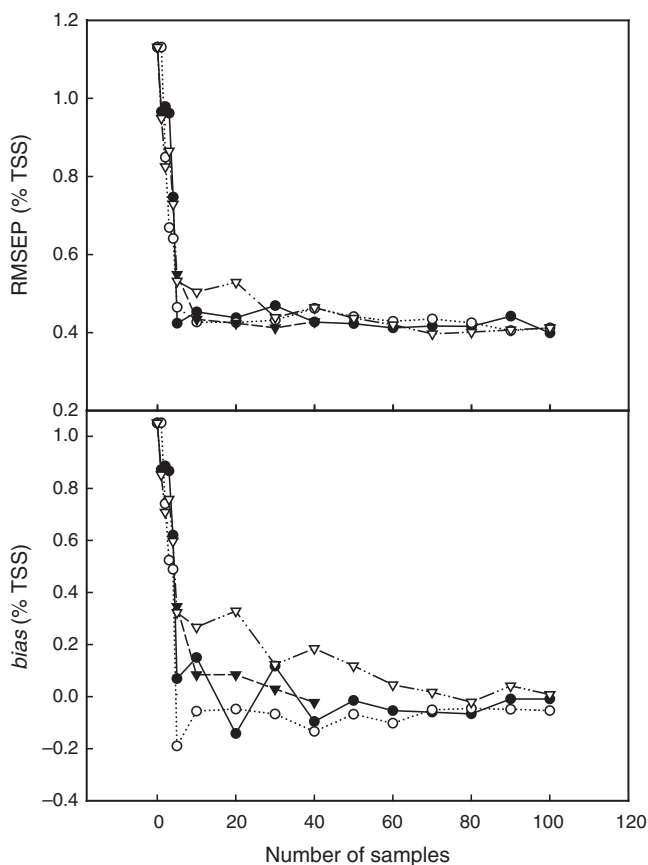


Fig. 1. Prediction statistics (RMSEP and *bias*) for a MPLS calibration model (using 2 populations from each of years 1999 and 2000) for mandarin fruit TSS. The independent validation population (another population from year 2000) was divided into 2 equal sets (1 set used for validation and the other used for sample addition to the calibration population). Four methods for sample selection for addition to the calibration population were used: (a) samples chosen randomly, closed circle; (b) every ‘*ith*’ sample based on ranking by GH, closed triangle; (c) samples with greatest GH less than 3, open circle; and (d) samples with the greatest NH, open triangle.

and *bias*) by inclusion of samples from the independent population, regardless of the method used to select the samples for inclusion (Fig. 3). The 3 methods of sample selection used (random, every ‘*ith*’ sample based on ranking by GH, and maximum GH) all behaved similarly, reaching a stable value after the addition of 10 samples.

The effect of sample addition (using the random selection method) on the performance of a TSS model (as used for Fig. 2; based on Populations J and K) was described for a further 3 independent validation populations (Populations E, G, and L). A similar activity was undertaken for DM (calibration Population T and validation Populations V, R, and S). The GH of the validation population was calculated using scores and loadings of the calibration population, with recalculation after each sample addition. Where the average GH of the validation population was markedly

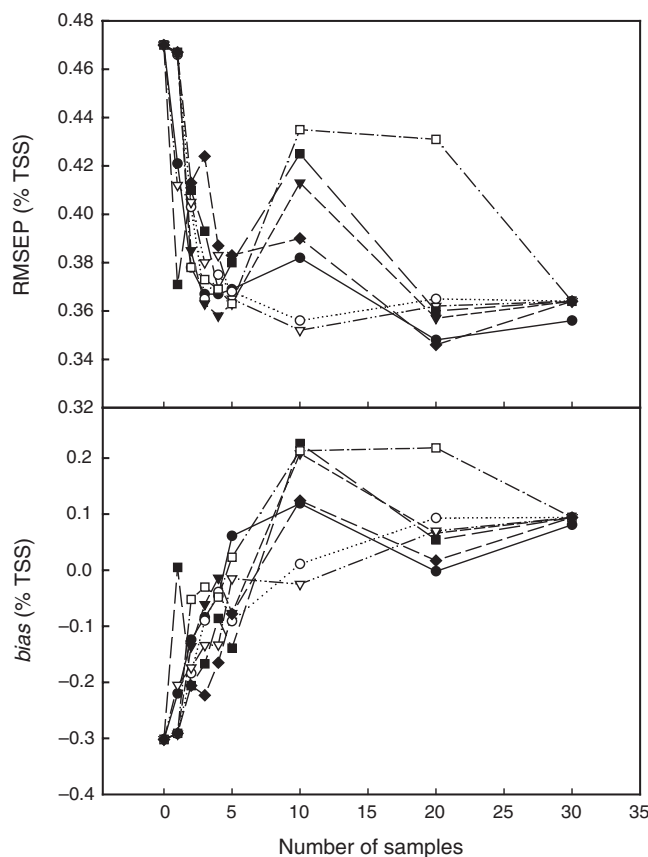


Fig. 2. Prediction statistics (RMSEP and *bias*) for a MPLS calibration model (Populations J and K) for mandarin fruit TSS. The independent validation population (Population M) was divided into 2 sets, two-thirds used for validation and the remainder used for sample addition to the calibration population. Six methods for sample selection for addition to the calibration population were used: (a) samples chosen randomly, open triangle and closed square; (b) samples with minimum GH values, closed circle; (c) samples with greatest GH less than 3, closed triangle; (d) every ‘*ith*’ sample based on ranking by GH, open circle; (e) samples selected using WINISI sample addition facility, open square; and (f) samples with the greatest NH, closed diamond.

different from the calibration population (e.g. average GH >3), the improvement in validation was quite dramatic (e.g. RMSEP decreasing from 1.45 to <0.60% TSS with the addition of only 5 samples). When the average GH of the validation population was similar to the calibration population (i.e. GH <3), the validation performance, although initially acceptable, showed little improvement (e.g. RMSEP changed from 0.50 to 0.42% TSS with addition of 5 samples for a population with an initial average GH of 2) (Figs 3 and 4).

Sample temperature

Model statistics (RMSEP) for DM prediction were not significantly different for calibration models developed using spectra of fruit at either 10 or 22°C, but that for 30°C was inferior to that at 22°C (Table 2). For calibrations

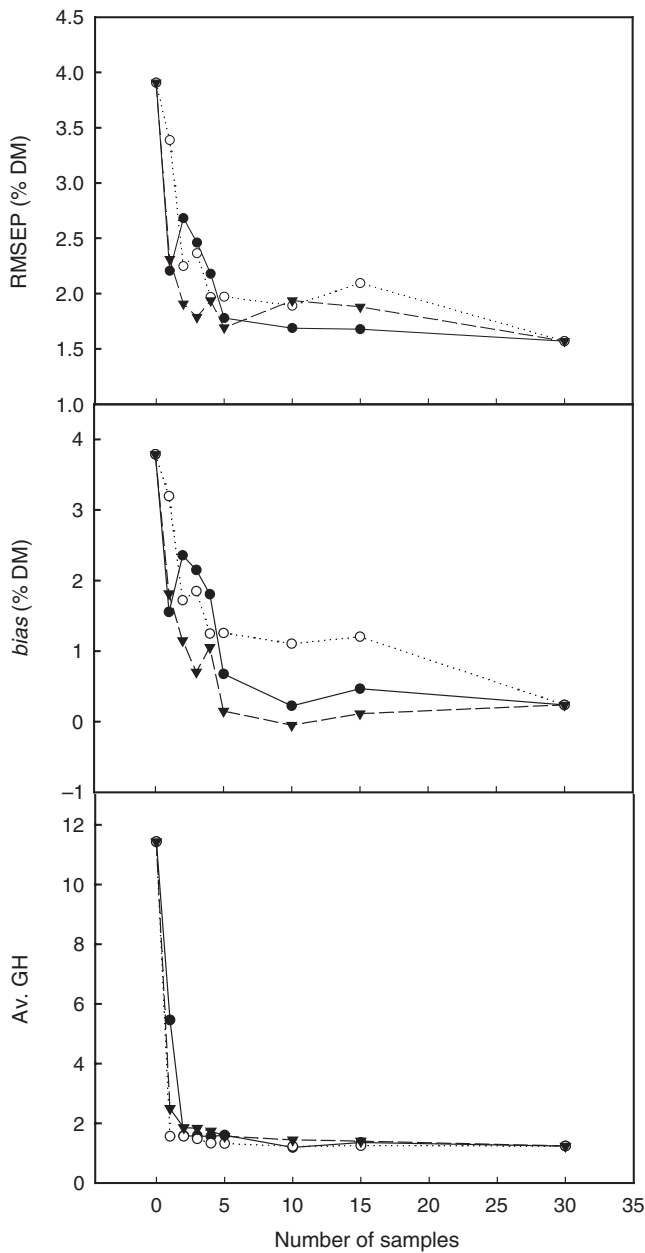


Fig. 3. Prediction statistics (RMSEP and *bias*) for MPLS prediction models for DM of mandarin fruit. The independent validation population (Population V) was divided into 2 sets, two-thirds used for validation and the remainder used for sample addition to the calibration population (Population T). Three methods for sample selection for addition to the calibration population were used: (a) samples chosen randomly, closed triangle; (b) samples with every 'ith' sample based on ranking by GH, closed circle; and (c) samples with greatest GH less than 3, open circle. The average GH of samples in the validation population was calculated using calibration model scores.

developed on TSS for these 3 fruit temperatures, calibration model RMSEP was not significantly different for models developed at either 10 or 22°C, but a significantly lower RMSEP was achieved at 30°C, compared with that at 22°C

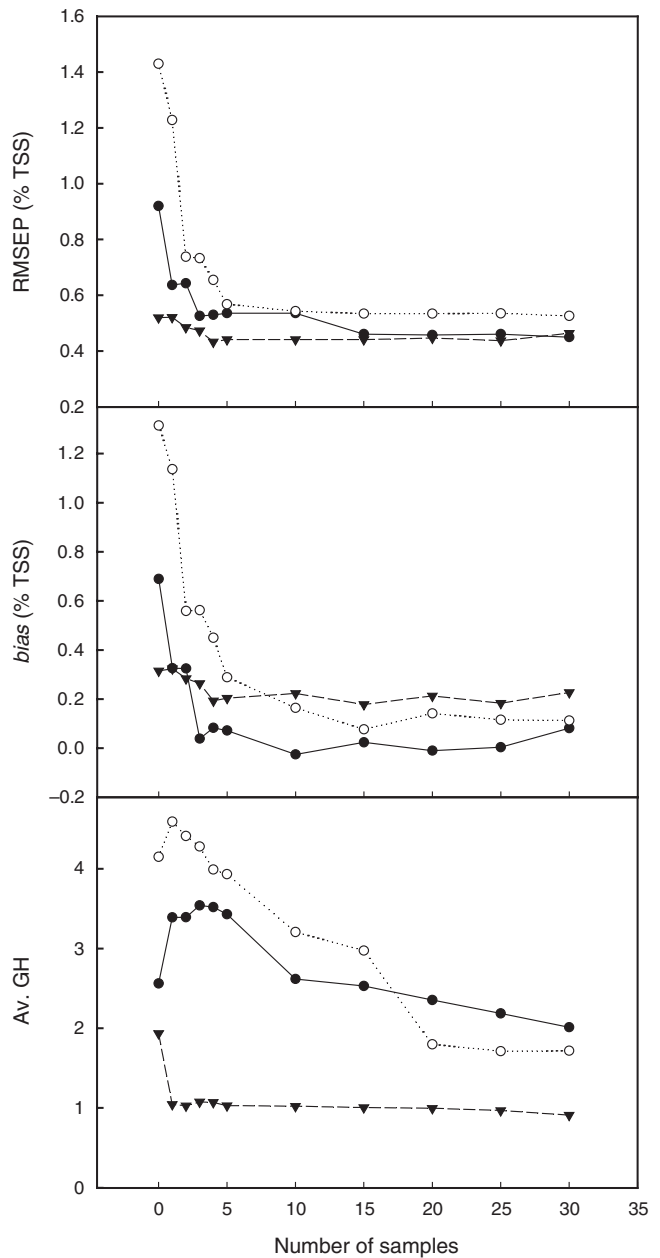


Fig. 4. Prediction statistics (RMSEP and *bias*) for MPLS prediction models for TSS of mandarin fruit, using 3 independent (of the calibration Populations J and K) validation populations (Populations E, closed circle; G, open circle; and L, closed triangle). The average GH of samples in the validation population was calculated using calibration model scores. Samples were selected randomly from the validation populations for addition to the calibration population.

(Table 2). Relative to models developed using fruit at several temperatures, a model (TSS or DM) developed at a single temperature (22°C) produced an inferior result [in terms of *bias* rather than RMSEP(C)] when fruit temperatures were other than that of the calibration population. For both the attributes of TSS and DM, *bias* was related to

fruit temperature ($R^2 = 0.96$, slope = -0.10% TSS/ $^{\circ}\text{C}$, and $R^2 = 0.86$, slope = -0.10% DM/ $^{\circ}\text{C}$).

Incorporating samples with different temperatures in the calibration population improved the prediction performance of both the TSS and DM models, in prediction of samples within the temperature range included in the calibration population (Table 2). For example, *bias* was -1.17 , -0.15 , and 0.07 for TSS models developed at sample temperatures of 22°C only, 10 and 22°C , and 10, 22, and 30°C , respectively, for prediction of a population of samples at 30°C .

The 4 samples with the lowest GH value from the calibration population scanned at 22°C were identified. The spectra of these samples at all 3 temperatures were included in the repeatability file of WINISI. In a second exercise, all samples (from across all temperatures) were used in the repeatability file. Including spectra of fruit scanned at different temperatures in the repeatability file did not improve calibration model statistics (for either DM or TSS), or model prediction statistics for DM, relative to a model using fruit re-scanned at all temperatures (Table 2). In contrast, the repeatability file option supported better prediction statistics for TSS, in terms of both RMSEP and *bias*, relative to a calibration developed using fruit re-scanned at all temperatures. Using all samples in the repeatability file was, however, better than using only 4 from each scanning temperature in this WINISI option.

Discussion

Calibration robustness: across seasons, locations, and harvest time

Validation of a model on a population independent of that used in calibration effectively tests for over-fitting of the model. Where the calibration model has weighted spectral features that represent fruit characteristics that are correlated to the attribute of interest in the calibration population, but not in the validation population, then validation performance will be poor. An example is a calibration developed for a variety in which skin chlorophyll content (skin greenness) is related to fruit TSS at maturity, which will not predict well with a variety in which there is no such relationship (unless the wavelength range considered is trimmed to eliminate the spurious correlation).

Calibration performance across harvest days (fruit from 1 tree in the 1 season) was superior to that across locations (fruit from harvests from varying farms in 1 season) (e.g. S/N statistic of 20 and 13, respectively, with an average RMSEP of 0.49 and 0.78% TSS), but performance was dramatically degraded when applied across seasons (S/N of 4, average RMSEP of 3.64% TSS). There was no trend for performance to degrade with increasing time (days) or distance/soil type of harvest (data not shown).

Taguchi descriptors calculated from 3 literature reports differ from those reported here. A S/N statistic of between

Table 2. Effect of temperature on prediction of DM and TSS for mandarin fruit

Models were developed on a population of 70 samples (SD 1.51% for TSS and SD 1.66% for DM), from spectra collected of these fruit at 10, 22, and 30°C . The prediction populations were based on a population of 34 samples (SD 1.64% for TSS and SD 2.03% for DM), again with spectra collected of these fruit at 10, 22, and 30°C . The calibration models for TSS involved 5 terms, whereas those for DM involved 6 or 7 terms. Separate calibration models were developed on spectra of fruit at 10, 22, 30, 10, and 22°C , and all 3 temperatures combined (i.e. 5 models). The WINISI repeatability file option was also used, with all samples or the 4 samples with lowest GH values from the 22°C validation population. The significance of the RMSEP at each temperature was tested. The RMSEP values followed by a common letter are not significantly different ($P < 0.05$). R_c^2 , coefficient of determination on calibration data set

Sample temperature ($^{\circ}\text{C}$)	Calibration model statistics				Prediction model statistics					
	<i>n</i>	RMSEP	R_c^2	RMSECV	<i>bias</i>	10 $^{\circ}\text{C}$ RMSEP(C)	<i>bias</i>	22 $^{\circ}\text{C}$ RMSEP(C)	<i>bias</i>	30 $^{\circ}\text{C}$ RMSEP(C)
<i>DM</i>										
10	70	0.63ab	0.85	0.71						
22	70	0.41a	0.94	0.51	0.54	0.79	0.04	0.52	-1.48	0.79
30	70	0.61b	0.86	0.66						
10 + 22	140	0.53	0.90	0.60	-0.26	0.77	-0.01	0.65	-0.78	0.81
10 + 22 + 30	210	0.55	0.89	0.60	-0.20	0.67	-0.01	0.56	-0.15	0.68
22 + repeatability	70	0.57	0.88	0.60	-0.11	0.84	-0.08	0.80	-0.14	0.86
<i>TSS</i>										
10	70	0.69a	0.79	0.75						
22	70	0.73a	0.77	0.88	0.89	1.16	0.22	1.13	-1.17	1.18
30	70	0.63b	0.82	0.76						
10 + 22	140	0.69	0.79	0.62	-0.08	1.11	0.17	1.11	-0.15	1.10
10 + 22 + 30	210	0.68	0.79	0.62	-0.05	1.14	0.38	1.19	0.07	1.17
22 + repeatability	70	0.69	0.79	0.81	-0.07	0.90	0.07	0.95	-0.07	0.89
22 + repeatability (4 samples)	70	0.79	0.73	0.89	-0.15	1.09	0.16	1.13	0.03	1.12

15 and 19, with an average RMSEP of approximately 1.1% TSS was calculated from the results of Peiris *et al.* (1998) for the use of peach TSS calibrations across 3 seasons. The S/N statistic and average RMSEP for the use of a single variety calibration model across other varieties was between 12 and 17, with an average RMSEP of approximately 1.0% TSS. The mandarin TSS predictions of Miyamoto and Kitano (1995) and Ou *et al.* (1997) yield a S/N statistic of 20 and average RMSEP of 0.58% TSS for predictions applied across seasons, and 7 (S/N) and 0.81% TSS (average RMSEP) across locations. Thus, previous studies indicate that model performance should be more stable across seasons, for a given variety, than across varieties, in a given season.

The cause of the dramatic decrease in performance of a calibration when applied to fruit across seasons in this study is not clear and could reflect changes in the instrument used as well as changes in the sample (fruit). However, there were no obvious changes in lamp or detector characteristics (i.e. in white reference spectra collected across years). The change in calibration performance between seasons is therefore more likely to represent changes in fruit optics (e.g. cell size, porosity), with consequent changes in the volume of fruit optically sampled, or in fruit composition (with characters other than the character of interest varying, and absorbing in similar wavelength regions).

Sample addition for calibration

To improve calibration performance on a new validation population, a common strategy is the addition of samples from the new population to the calibration population. The RMSEP and *bias* (Figs 1–5) decreased with addition of validation samples to the calibration population reaching a stable value with the addition of about 20 samples. Several approaches were used in the selection of samples from the validation population for addition to the calibration population; however (surprisingly), all methods performed equally well. This result indicates that the variation within a new population must be small, relative to the difference of that population to the calibration population, such that any sample chosen from within a given population is a useful representative of that population.

The higher the average GH of the validation population when calculated on the calibration population scores (Figs 4 and 5 for the attributes of TSS and DM, respectively), the greater the improvement to RMSEP and *bias* with the addition of validation population samples. Higher average GH values reflect an increased difference in the spectra of calibration and validation populations, and a greater leverage on the MPLS regression will be gained in sample addition from the validation population.

It is surprising, however, that the reverse was not true, i.e. that the addition of high GH validation samples to the calibration population was not more effective than

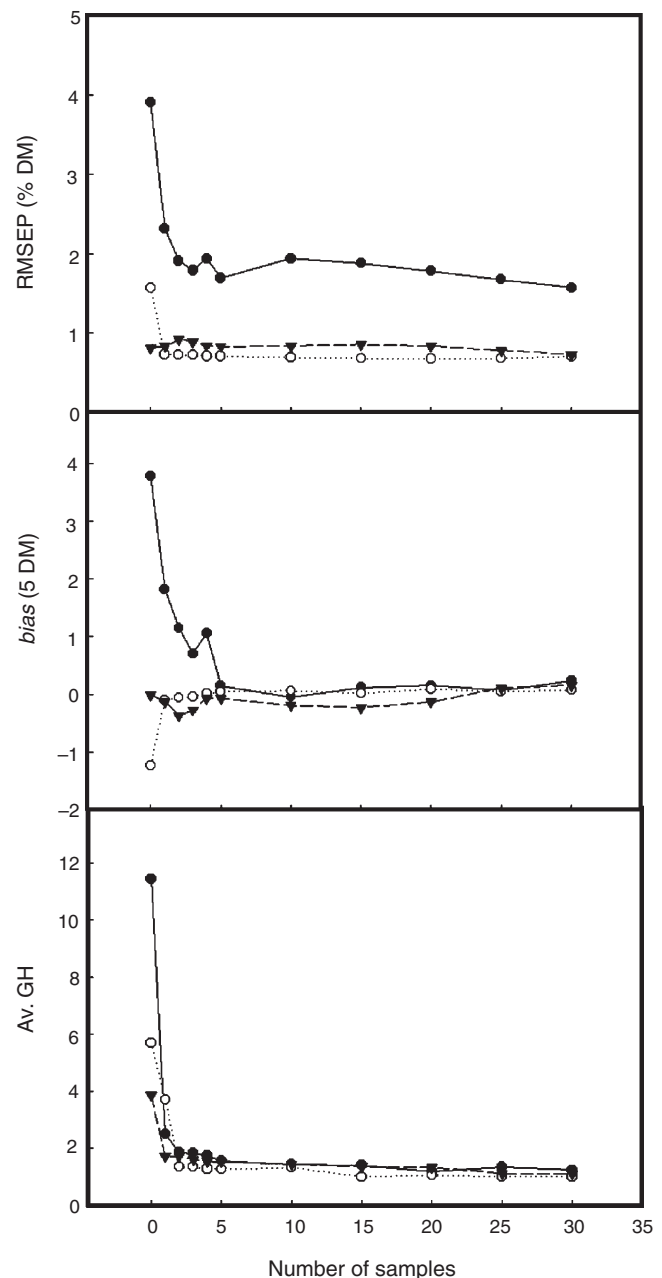


Fig. 5. Prediction statistics (RMSEP and *bias*) for MPLS prediction models for DM of mandarin fruit, using 3 independent (of the calibration population T) validation populations (Populations V, closed circle; R, open circle; and S, closed triangle). The average GH of samples in the validation population was calculated using calibration model scores. Samples were selected randomly from the validation populations for addition to the calibration population.

the addition of low GH validation samples, in terms of improvement to prediction RMSEP and *bias*.

In practice, the level of accuracy required must be established for each sorting task. Higher accuracy requirements will require higher calibration maintenance.

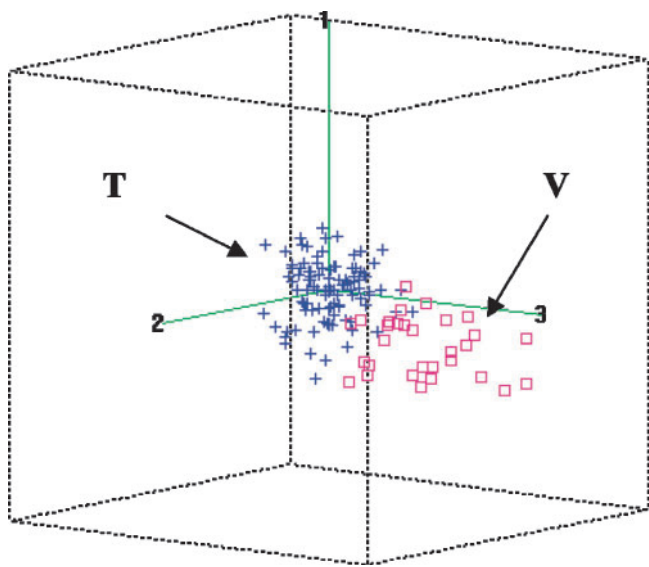


Fig. 6. Three-dimensional plot of MPLS scores (1, 2, and 3) of the calibration population T ($n=103$) and an independent set (subset of population V, $n=30$) calculated using a calibration model for DM.

This maintenance might involve adjustment of *bias* for new populations, or recalibration with addition of spectra of at least 20 fruit from the new population to the existing calibration population, to recover RMSEP values.

It is not obvious why the inclusion of such a small number of samples to the calibration population can have such an influential effect. It may be partly due to the added samples being so different (separate) from the original calibration population, resulting in 2 ‘clusters’ (original and new) (Fig. 6) that are basically treated by the calibration as 2 ‘points’. However, it is then intriguing that the model predicts so well on the validation population. Regardless, the methodology was observed to work well in a number of circumstances and for a number of populations.

Calibration robustness: temperature

Mandarin fruit temperature can vary from 5°C (recommended storage temperature) to over 30°C (field temperature) during in-line grading in a commercial packing shed. Temperature affects the degree of H bonding, and thus the position and intensity of OH stretching vibration bands. There are 2 main forms of liquid water, 1 form involving a H bond to another water molecule, and the other form involving more structured water. The second form dominates at lower temperatures, and absorbs at higher wavelengths relative to the first form. Golic *et al.* (2003) reported that calibration model statistics for models developed for pure sucrose solutions across a range of sample temperatures were degraded relative to those at a constant temperature (20°C). These calibrations resulted in a de-emphasis on those areas of the spectrum associated with OH stretching,

favouring those areas associated with other spectral bands of the sugars (e.g. 910 nm CH third overtone).

Where a model was required to predict samples with temperatures outside the range included in the calibration population, *bias* was increased for both DM and TSS models (Table 2). The RMSEP was affected primarily through an effect on prediction *bias*. Therefore the following discussion reports on *bias* and RMSEP(C). In practical application, a *bias* adjustment could be applied for the use of a calibration at temperatures outside of the range included in the calibration population.

Calibrations developed across a wide range of temperatures are expected to be more robust in terms of predicting analyte levels of samples at a range of temperatures, although potentially at the expense of diminished accuracy. Prediction robustness in terms of *bias* was indeed increased for models developed across several temperatures, whereas accuracy [RMSEP(C)] was similar to that of single temperature calibration models, for both DM and TSS (Table 2). Kawano *et al.* (1995) also found that incorporation of samples across a temperature range in a (MLR) calibration allowed prediction of TSS with a high degree of accuracy and minimal *bias*.

We expected DM calibration models to be more sensitive to temperature than TSS models, given the sensitivity of the water bands to temperature (H bonding status) (Golic *et al.* 2003). This was not so, with TSS and DM similarly sensitive to temperature (Table 2). Presumably this effect reflects the large contribution of sugar OH features that are sensitive to H bonding status, and thus to temperature, in both the TSS and DM calibration models.

The repeatability file option in WINISI was implemented in an attempt to reduce the sensitivity of the calibration to sample temperature variations. Wavelengths with less change due to temperature should receive higher PLS scores, thus decreasing emphasis on the remaining wavelengths. For TSS, the repeatability file option did not improve calibration statistics, but prediction was improved relative to a model incorporating spectra of fruit at all 3 temperatures. For DM, implementation, to the extreme of including all spectra in the repeatability file, was not as successful as the combined file approach.

Conclusions

Calibration models were less robust across seasons than across locations and time within a harvest season. In all cases, model updating involving the addition of relatively few samples (approx. 20) was successful in improving prediction of new populations. The method of sample addition was not crucial. Therefore, for ease of operation the random selection approach is the logical choice for sample addition to improve RMSEP and *bias* in the prediction of independent validation populations (for both the attributes of TSS and

DM). The higher the average GH of the independent population with respect to the scores and loadings of the calibration population, the greater the beneficial effect of sample addition.

We conclude, in agreement with Miyamoto and Kitano (1995) and (Kawano *et al.* 1995), that samples scanned at a range of temperatures should be included in the calibration population in order for the model to be robust in prediction of samples varying in sample temperature. The issue of calibration population design to incorporate robustness for sample temperature without loss of general validation accuracy (i.e. what proportion of calibration samples should be run at different temperatures, and over what number of temperature steps) requires further consideration. Alternatively, the orthogonal projection method suggested recently by Roger *et al.* (2003) in a consideration of model robustness across instruments may have merit for increasing calibration robustness to temperature variation.

Acknowledgments

Funding support was received from Horticulture Australia (Citrus Marketing and Development Group) and Central Queensland University – Research Training Scheme. Fruit was supplied by Steve Benham of Joey Citrus, Munduberra, and by Jim and Deslea Yeldham, Citrus Farm, Dululu, Qld. This manuscript represents an extension of work reported at the 10th International Conference of Near Infrared Spectroscopy, Konju, Korea, June 2001.

References

- Fearn T (1996) Comparing standard deviations. *NIR News* **7**, 5–6.
- Golic M, Walsh KB, Lawson P (2003) Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature. *Applied Spectroscopy* **57**, 139–145. doi: 10.1366/000370203321535033
- Greensill CV, Walsh KB (2000) A remote acceptance probe and illumination configuration for spectral assessment of internal attributes of intact fruit. *Measurement Science & Technology* **11**, 1674–1684. doi: 10.1088/0957-0233/11/12/304
- Greensill CV, Walsh KB (2002) Calibration transfer between miniature photodiode array-based spectrometers in the near infrared assessment of mandarin soluble solids content. *Journal of Near Infrared Spectroscopy* **10**, 27–35.
- Guthrie JA, Walsh KB (2001) Assessing and enhancing near infrared calibration robustness for soluble solids content in mandarin fruit. In '10th International Conference on Near Infrared Spectroscopy'. Kyonjgu, Korea. (Eds AMC Davies, RK Cho) pp. 151–154. (NIR Publications: Chichester, UK)
- Guthrie JA, Walsh KB, Reid DJ, Liebenberg CJ (2005) Assessment of internal quality attributes of mandarin fruit. 1. NIR calibration model development. *Australian Journal of Agricultural Research* **56**, 405–416.
- Kawano S, Abe H, Iwamoto M (1995) Development of a calibration equation with temperature compensation for determining the brix value in intact peaches. *Journal of Near Infrared Spectroscopy* **3**, 211–218.
- Mahalanobis PC (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* **2**, 541–588.
- McGlone VA, Fraser DG, Jordan RB, Kunnemeyer R (2003) Internal quality assessment of mandarin fruit by vis/NIR spectroscopy. *Journal of Near Infrared Spectroscopy* **11**, 323–332.
- Miyamoto K, Kitano Y (1995) Non-destructive determination of sugar content in satsuma mandarin fruit by near infrared transmittance spectroscopy. *Journal of Near Infrared Spectroscopy* **3**, 227–237.
- Ou AS, Lin S, Lin T, Wu S, Tiarn M (1997) Studies on the determination of quality-related constituents in Ponkan Mandarin by near infrared spectroscopy. *Journal of the Chinese Agricultural Chemical Society* **35**, 462–474.
- Peiris KHS, Dull GG, Leffler RG, Kays SJ (1998) Near-infrared spectroscopic method for non-destructive determination of soluble solids content of peaches. *Journal of the American Society for Horticultural Science* **123**, 898–905.
- Roger J-M, Chauchard F, Bellon-Maurel V (2003) EPO and PLS external parameter orthogonalisation of PLS—application to temperature independent measurement of sugar content of intact fruits. *Chemometric and Intelligent Laboratory Systems* **66**, 191–204.
- Sanchez HN, Luro S, Roger JM, Bellon-Maurel V (2003) Robustness of models based on NIR spectra for sugar content prediction in apples. *Journal of Near Infrared Spectroscopy* **11**, 97–107.
- Shenk JS, Westerhaus MO (1991) Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. *Crop Science* **31**, 469–474.
- Wang Y, Veltkamp DJ, Kowalski BR (1991) Multivariate instrument standardization. *Analytical Chemistry* **63**, 2750–2756. doi: 10.1021/ac00023a016
- Westerhaus MO (1991) Improving repeatability of calibrations across instruments. In 'Proceedings of 3rd International Conference on Near Infrared Spectroscopy'. Gembloux, Belgium. (Eds R Biston, N Bartiaux-Thill) pp. 671–674. (NIR Publications: Chichester, UK)
- Williams PC, Sobering DC (1993) Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy* **1**, 25–32.
- Wortel VAL, Hansen WG, Wiedemann SCC (2001) Optimising multivariate calibration by robustness criteria. *Journal of Near Infrared Spectroscopy* **9**, 141–151.

Manuscript received 2 December 2004, accepted 23 February 2005