Wavelength selection for partial least squares-based visible–near infrared evaluation of soluble solids content in fresh fruits

Paolo Carlini,* Riccardo Massantini and Fabio Mencarelli

Istituto di Tecnologie Agroalimentari, Facoltà di Agraria, Università degli Studi della Tuscia, Via S.C. De Lellis snc, I-01100, Viterbo, Italy.

Introduction

There are few independent research groups investigating the actual realisation of complete visible–near infrared (Vis–NIR) systems able to grade and sort fruits and vegetables on-line.[1–3] Current research is aimed mostly at an accurate assessment of the attainable analytical accuracy, for many different products and many different chemical constituents and physicochemical properties (e.g. firmness).[4–7]

The experiments reported here were primarily concerned with Vis–NIR interactance measurement of soluble solids content (SSC) on intact fruits, highly correlated to total sugar content. This parameter, only one of the many jointly defining the "internal" quality of a fresh fruit is, however, one of the most relevant in the study of ripening processes (see, for example Reference 8).

We believe that the general strategy adopted, involving the use of *many* disjointed spectral segments in a partial least squares (PLS) model (v. one big interval, as usually reported in the literature) should be potentially useful for other constituents too, such as titratable acidity, starches, simple sugars and others.

Some species studied—apricot and loquat fruit—have never been considered before for Vis–NIR non-destructive quality determination, with or without the use of modern regression techniques or wavelength selection methods.

Materials and methods

Fruits were either hand-harvested at the beginning of summer, 1998, from a small orchard in Vetralla (Viterbo, Italy)—peach (*Prunus persica* L. cv. "Royal moon")—or purchased in the market in spring/summer, 1997—apricot (*Prunus armeniaca* L. cv. "Boccuccia Spinosa" and "Errani") and loquat fruit [*Eriobotrya japonica* (Thunb.) Lindley]—selecting first class samples, uniform in size, then immediately brought to our laboratory and evaluated at room temperature.

The absorption spectrum was measured on each intact sample using a Vis–NIR spectrophotometer NIRSystems (Silver Spring, MD, USA) model 6500 equipped with a fibre-optic probe, working by interactance. An outer ring, about 1 cm in diameter, emitted the light interacting with the sample, a central fibre bundle returned it. For system management and calibration NIRS-*2* Version 4.00 package by Infrasoft International (Port Matilda, PA, USA) was adopted. Spectra were measured by hand-placing the probe against the fruit at a random position along the equator. Fifteen individual scans were averaged for the recording of each spectrum.

Near Infrared Spectroscopy: Proceedings of the 9th International Conference

Reference SSC readings have been taken for flesh belonging to the location of the Vis–NIR measurement. Pulp was slightly comminuted and centrifuged for five minutes by an ALC micro centrifugette 4204. The surnatant was then analysed by a laboratory refractometer Officine Galileo (Florence, Italy) model RG701.

Prior to model building, a randomised procedure split each spectrum/reference SSC data set into a *calibration* set and a smaller size *prediction* set, used for validation on independent samples. A slightly modified form of PLS was used as the multivariate regression algorithm, involving the normalisation of the residuals at the end of each iteration. Calibrations have only been considered if their complexity, in terms of number of PLS factors, was below a maximum order, established case by case by using an 8-way cross-validation strategy.

Well-suited intervals have been found by way of trial and error procedures in every instance, including a large segment in the vis–Herschel-Infrared to begin with. Usually, a suitable set of pre-processing options could be established early in the work and remain fixed during the second phase, whereas models were improved by purging some non-informative intervals and adding a few narrow segments in the proper NIR region.

The latter have been defined one after the other, wavelengths closest to the vis–Herschel region the first to be screened, in a sort of *forward selection* procedure[9] guided by *SEP* statistics and generalised to whole segments. A final refinement stage reconsidered and slightly altered the extremes of some intervals. For each spectral segment a sampling step had also to be chosen, which was a multiple of 2 nm.

Results and discussion

Table 1 shows the soluble solids levels in calibration and prediction apricots, loquat fruits and peaches.

Table 1. Soluble solids level (°Brix) in calibration and prediction apricots, loquat fruits and peaches.

	Samples	Mean	Std dev.	Min.	Max.
Apricot					
Calibration	120	11.33	1.91	7.45	16.50
Prediction	42	11.61	2.07	7.80	17.00
Total	162	11.40	1.95	7.45	17.00
Loquat fruit					
Calibration	70	10.05	1.46	7.30	13.73
Prediction	40	10.35	1.61	7.35	13.73
Total	110	10.16	1.51	7.30	13.73
Peach					
Calibration	66	14.66	1.65	11.85	19.52
Prediction	42	14.58	1.64	11.60	17.95
Total	108	14.63	1,64	11.60	19.52

Table 2. Calibration statistics of models for apricot. *SEC*, *SEP* and Bias in °Brix.

Full spectrum					
PLS Factors	*SEC*	R^2	*SEP*	Bias	*SDR*
6	0.72	0.86	0,84	0.05	2.32
7	0.65	0.88	0.81	0.10	2.41
8	0.58	0.91	0.73	0.03	2.68
9	0.52	0.92	0.66	0.00	2.93
Wavelength selection					
PLS Factors	*SEC*	R^2	*SEP*	Bias	*SDR*
7	0.63	0.89	0.71	0.09	2.74
8	0.56	0.91	0.61	0.04	3.17
9	0.46	0.94	0.43	0.03	4.51
10	0.43	0.95	0.42	0.01	4.63

Apricot

Shortly after data taking, a preliminary full spectrum model, involving the spectral region between 650 and 2060 nm, step 4 nm, was established, which can be usefully contrasted to recent results. The software package had selected a maximum number of factors equal to nine and the resulting predictive accuracy had been not very satisfying, with *SEP* about 20% larger than *SEC* (Table 2). Figure 1(a) shows the prediction set scatter plot corresponding to nine factors and a sensible spread around the bisecting line can be clearly appreciated.

Starting from scratch, far better performing models have been found recently, involving the following spectral intervals:
662–846, 4; 854–880, 4; 888–1088, 4; 1108–1210, 4;
1220–1300, 4; 1320–1454, 4; 1460–1520, 4; 1540–1590, 4; 1624–1686, 4; 1720–1740, 4;
1758–1758, 2; 1810–1890, 4; 1920–1970, 4; 1980–2064, 4; 2162 – 2166, 2.

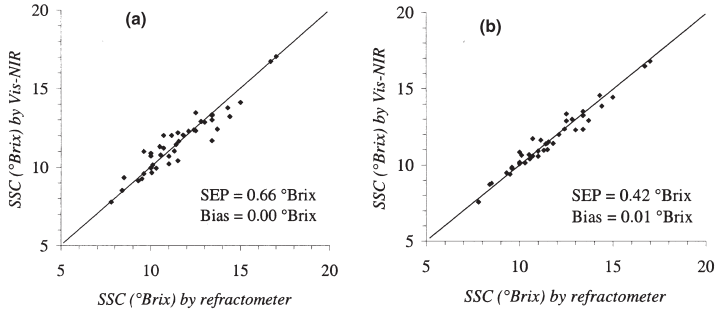

Figure 1. Scatter plot for prediction of apricots: (a) full spectrum; nine PLS model points v. (b) ten PLS factor model exploiting the wavelength selection.

Apart from a single line at 1758 nm, another at 2964 nm and a triplet between 2162 and 2166 nm, only wavelengths belonging to a *subset* of those included in the full spectrum calibrations above were included in the new models.

Performance statistics, shown in the second part of Table 2 for a number of PLS factors ranging from 7 to 10, were much improved in every possible respect.

Let us remark that every pre-processing option was, in fact, the *same* as for the preliminary models: I derivative, six point *gap* in the numerical computation of the derivative and three point smoothing, weighted MSC scattering correction. The only news was the detection of two outliers. Figure 1(b) shows the prediction set scatter plot for ten PLS factors.

Loquat fruit

For loquat fruit two families of models have been found, invariably utilising a II-derivative pre-treatment computed with a 19 point gap and preceded by a nine point smoothing pass, SNV and Detrend scattering correction and involving four and nine intervals, respectively:
600–1050,8; 1320–1390,8; 1500–1790,8; 2130–2200,8;
600–720, 8; 730–760, 8; 770–1050, 8; 1320–1390, 8; 1500–1790, 8; 1900–1940, 8; 2110–2130, 8; 2140–2160, 8; 2170–2200, 8.

Table 3 shows regression statistics for a number of factors ranging from 7 to 11.

Peach

Regarding this species, the state of the art was represented by two known studies.[10,6] Our tests confirmed that peach is quite difficult to deal with and only after much effort have we been able to establish decent models, involving many different step values, in the range from 2 to 8 nm and a few individual wavelengths. For an I-derivative pre-processing computed with a six point gap and no smoothing, weighted MSC scattering correction, the optimal wavelength intervals found are as follows:

Table 3. Calibration statistics of models for loquat fruit. *SEC*, *SEP* and Bias in °Brix.

4 intervals					
PLS factors	*SEC*	R^2	*SEP*	Bias	*SDR*
7	0.51	0.88	0.54	0.04	2.82
8	0.48	0.89	0.52	0.03	2.91
9	0.43	0.91	0.50	0.01	3.03
10	0.40	0.92	0.45	0.00	3.39
11	0.33	0.95	0.37	–0.00	4.12
9 intervals					
PLS factors	*SEC*	R^2	*SEP*	Bias	*SDR*
7	0.51	0.87	0.56	0.02	2.71
8	0.47	0.89	0.51	0.03	2.94
9	0.43	0.91	0.44	0.02	3.40
10	0.35	0.94	0.36	–0.04	4.18
11	0.32	0.95	0.34	–0.01	4.46

Table 4. Calibration statistics of models for peach. *SEC*, *SEP* and Bias in °Brix.

PLS factors	*SEC*	R^2	*SEP*	Bias	*SDR*
6	0.75	0.80	0.72	–0.00	2.26
7	0.57	0.88	0.48	–0.10	3.43
8	0.49	0.91	0.45	–0.01	3.66
9	0.47	0.92	0.40	–0.02	4.14

520–536, 8; 572–608, 8; 620–654, 8; 688–780, 4; 804–1098, 4;
1108–1130, 4; 1172–1178, 2; 1296–1296, 2; 1398–1398, 2; 1440;
1516; 1666; 1898–1914, 4; 2104–2110, 2; 2120–2128, 4; 2150–2160, 4; 2244–2252, 4; 2258; 2302.

A remarkable feature of the models is a sensible difference between *SEC* and *SEP* values, the latter being much lower than the former (Table 4). Probably, a data set enlargement may help to diagnose some difficulties experienced on peaches better.

Conclusions

In this short communication we have brought preliminary results favourable to the use of wavelength selection for PLS-based evaluation of SSC in fruits, extending the existing body of knowledge, to date restricted to the kiwifruit case.[11]

Of course, much work remains to be done, on more species and many varieties at the same time, which could also be helpful in developing new, effective techniques for the automatic or semi-automatic selection of suited intervals.

References

1. S. Kawano, *Jpn Agric. Res. Q.* **28,** 212 (1994).
2. J.M. Roger, V. Bellon-Maurel, L. Dusserre-Bresson, P. Fayolle and G. Ranou, in *SENSORAL 98: Sensing Quality of Agricultural Products*. Montpellier, France, pp. 24–27 (1998).
3. Z. Schmilovitch, A. Hoffman, H. Egozi, R. Ben-Zvi, Z. Bernstein and V. Alchanatis, *J. Sci. Food Agric.* **79,** 86 (1999).
4. V.A. McGlone and S. Kawano, *Postharvest Biol. Technol.* **13,** 131 (1998).
5. J. Lammertyn, B. Nicolaï, K. Ooms, V. De Smedt and J. De Baerdemaeker, *Trans. ASAE* **41,** 1089 (1998).
6. K.H.S. Peiris, G.G. Dull, R.G. Leffler and S.J. Kays, in *Sens. Nondestr. Test. Int. Conf.* Northeast Reg. Agric. Eng. Serv., Ithaca, New York, USA, p. 77 (1997).
7. K.H.S. Peiris, G.G. Dull, R.G. Leffler and S.J. Kays, *J. Amer. Soc. Hort. Sci.* **123,** 1089 (1998).
8. S.J. Kays, *Postharvest physiology of perishable plant products*. Van Nostrand Reinhold, New York, USA (1991).
9. A.J. Miller, *Subset Selection in Regression*. Chapman & Hall (1990).
10. D.C. Slaughter, *Trans. ASAE* **38,** 617 (1995).
11. S.D. Osborne, R.B. Jordan and R. Künnemeyer, *Analyst* **122,** 1531 (1997).