J Am Oil Chem Soc (2008) 85:1105–1113
DOI 10.1007/s11746-008-1311-1

ORIGINAL PAPER

Measurement of Whole Soybean Fatty Acids by Near Infrared Spectroscopy

Benoit Igne · Glen R. Rippke · Charles R. Hurburgh Jr.

Received: 30 May 2008 / Revised: 25 September 2008 / Accepted: 2 October 2008 / Published online: 1 November 2008

Abstract Whole soybean fatty acid contents were measured by near infrared spectroscopy. Three calibration algorithms—partial least squares (PLS), artificial neural networks (ANN), and least squares support vector machines (LS-SVM)—were implemented. Three different validation strategies using independent sets and part of calibration samples as validation sets were created. There was a significant improvement of the prediction precision of all fatty acids measured on relative concentration of oil compared with previous literature using PLS (standard error of prediction of 0.85, 0.42, 1.64, 1.67, and 0.90% for palmitic, stearic, oleic, linoleic and linolenic acids respectively). ANN and LS-SVM methods performed significantly better than PLS for palmitic, oleic and linolenic acids. Calibration models developed on relative concentrations (% of oil) were compared to prediction models created on absolute fatty acid concentration (% of weight) and corrected to relative concentration by multiplying by the predicted oil content. While models were easier to develop in absolute concentration (higher coefficients of determination), the multiplication of errors with the total oil content model resulted in no net precision improvement.

Keywords Partial least squares · Artificial neural networks · Least squares support vector machines · Near infrared spectroscopy · Fatty acids · Soybeans

B. Igne (✉) · G. R. Rippke · C. R. Hurburgh Jr.
Department of Agricultural and Biosystems Engineering, Iowa State University, 1551 Food Sciences Building, Ames, IA 50011, USA
e-mail: igneb@iastate.edu

Introduction

Modern society has developed food habits and industries that favor fast foods and high fat diets. Fat intake has been targeted as one of the most critical parameters to control for reducing the risk of cardiovascular diseases [1–4]. Linoleic and linolenic acids, essential poly-unsaturated fatty acids, lower the levels of low-density lipoprotein cholesterol in the blood. This type of cholesterol is thought to be responsible for accumulations in the arteries [5, 6]. Research points to a balanced intake of linoleic and linolenic acids to lower health risks [7]. Saturated fatty acids are also of interest for processing purposes. Palmitic and stearic acids have demonstrated useful properties for the production of margarine and shortening [8, 9].

Soybean and canola oil are the only edible oils that present a high level of both linoleic and linolenic acids [10]. Certain uses of these oils, e.g., for frying, requires hydrogenation, a process that adds hydrogen, but produces trans fatty acids which increase the risk of heart diseases [11]. Breeders have developed varieties of soybeans with altered fatty acid profiles to match various applications (low linolenic oils, low saturated oils) [12].

Near infrared spectroscopy (NIRS) is a fast, non-destructive, and inexpensive technique for routine utilization that has shown, for more than thirty years, its usefulness in the control of grain quality [13–16]. NIRS has been approved by the American Association of Cereal Chemists (AACC) for determining protein content of whole-grain wheat (Method 39-00) and by the American Oil Chemists' Society (AOCS) as a procedure providing general guidelines (Procedure Am 1–92). Several authors have published significant works on the prediction of fatty acids by near infrared spectroscopy. Velasco et al. [17]

reported the ability of NIRS to predict, in cross-validation, oleic, linoleic and linolenic acids for intact rapeseed against gas-liquid chromatography reference measurement ($r^2 = 0.98$, $r^2 = 0.95$, and $r^2 = 0.96$ respectively). Other studies reported the ability of NIRS to measure fatty acids in whole sunflowers achenes [18] and in peanut seeds [19]. In soybean, Pazdernik et al. [20] reported low correlations between NIRS and fatty acid reference measurements on whole grain (palmitic acid: $r^2 = 0.36$, standard error of cross validation (SECV) = 0.79%; stearic acid: $r^2 = 0.74$, SECV = 0.22%; oleic acid: $r^2 = 0.58$, SECV = 0.77%; linoleic acid: $r^2 = 0.61$, SECV = 1.71%; linolenic acid: $r^2 = 0.84$, SECV = 0.68%), but better results on ground samples (palmitic acid: $r^2 = 0.59$, SECV = 0.52%; stearic: $r^2 = 0.72$, SECV = 0.19%; oleic: $r^2 = 0.83$, SECV = 0.56%; linoleic acid: $r^2 = 0.89$, SECV = 0.91%; linolenic acid: $r^2 = 0.89$, SECV = 0.42%).

Lately, Nimaiyar et al. [21] published validation results on whole soybean samples using an independent validation set [palmitic acid: $r^2 = 0.40$, root mean standard error of prediction (RMSEP) = 0.79%; stearic acid: $r^2 = 0.24$, RMSEP = 0.39%; oleic acid: $r^2 = 0.59$, RMSEP = 3.46%; linoleic acid: $r^2 = 0.76$, RMSEP = 2.37%; linolenic acid: $r^2 = 0.84$, RMSEP = 0.55%]. These results show the lack of robustness of prediction models when applied to new samples. This study as well as the one of Pazdernik et al. was performed on a limited number of calibration samples (70 or less). In 2006, Kovalenko et al. [22] published a larger study of the prediction of fatty acids on whole soybean samples using more calibration samples (600 or more) and different regression techniques (linear and non-linear). They reported good correlations between NIRS data and fatty acids measured by gas chromatography for saturated acids (palmitic acid + stearic acid) [$r^2 = 0.91$, standard error of prediction (SEP) = 2.23%], palmitic acid ($r^2 = 0.80$, SEP = 3.16%), oleic acid ($r^2 = 0.76$, SEP = 4.27%), and linoleic acid ($r^2 = 0.73$, SEP = 3.77%), but a more limited relationship for stearic acid ($r^2 = 0.49$, SEP = 0.47%) and linolenic acid ($r^2 = 0.67$, SEP = 1.74%) using partial least squares (PLS) regression. Significantly better results were obtained with non-linear regression techniques (ANN and LS-SVM).

All the work cited above developed calibrations in relative concentrations of fatty acids (grams of fatty acid by 100 g of oil).

The objectives of this study were (1) to improve present calibration performance by including the variability in fatty acid composition present in the US market using both linear and non-linear regression methods and (2) to evaluate the possibility of developing fatty acid calibrations in absolute concentration (gram of fatty acid by 100 g of sample). Predicted absolute concentrations are converted to relative concentrations by dividing the predicted absolute concentration by the predicted oil content of the same sample.

Experimental Procedure

Samples, Spectra Collection and Reference Analysis

A set of approximately 900 whole US soybean samples (400 g per sample or more) from crop years 2003 to 2006 were scanned on four near infrared instruments with a pathlength of 30 mm. Foss Infratec Grain Analyzers 1229 and 1241 (FOSS North America, Eden Prairie, MN, USA) and two Bruins OmegAnalyzerG (serial: 106110 and 106118) (Bruins Instruments, Puchheim, Germany) were used. Both are transmittance units with a spectral range from 850 to 1,048 nm at an increment of 2 nm. These instruments were chosen because they are approved by the US National Type Evaluation Program (NTEP) [23–25] and are by consequence approved to perform analysis for trade in the USA. Samples were run at room temperature. Each sample was run simultaneously on the four instruments. Oil content was determined by ether extract (AOCS Method Ac 3-44) by Eurofins Scientifics, Inc., Des Moines, IA, USA and the relative concentrations of palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1), linoleic acid (C18:2) and linolenic acid (C18:3) were analyzed by gas chromatography using the method described by Hammond [26] in the Department of Agronomy at Iowa State University, Ames, IA, USA. The total saturated fatty acid concentration was determined by adding for each sample the relative concentrations of palmitic and stearic acids. Absolute fatty acid concentrations were calculated by multiplying relative concentrations expressed in percent of oil by the oil content of the sample, from lab chemistry in calibration situations and from predictions in validation situations. Summary statistics for the calibration and validation sets are presented in Table 1.

Spectral Pretreatment and Outlier Detection

Raw spectral data (log (1/T) vs. wavelength) were corrected for baseline and scattering effect by calculating their second derivative spectra using the Savitzky–Golay [27] algorithm (5-point window and 3rd order polynomial) and each sample was normalized to the sum of the absolute value of all variables (wavelength) for the given sample and instrument (Fig. 1). Furthermore, variables were scaled to zero mean and unit standard deviation. The detection of outliers was performed on the fully pretreated spectra by removing from the calibration set samples presenting a Hotelling T^2 and a Q residual value (eigenvalue of the residual subspace) larger than the 95% confidence interval.

Table 1 Reference data statistics for soybean samples used for calibration and validation

Parameters	2003–2005 samples[a]				2006 samples[a]			
	n[b]	Average concentrations[c]	Range	SD	n[b]	Average concentrations[c]	Range	SD
Saturated	785	12.49	5.61–18.93	4.12	126	14.58	6.33–17.45	2.20
Palmitic	732	8.16	2.89–13.64	3.45	115	10.34	3.40–13.47	1.86
Stearic	750	4.32	2.62–6.81	0.84	135	4.24	2.48–6.54	0.84
Oleic	801	26.02	19.42–36.91	2.89	161	25.72	20.51–36.84	3.64
Linoleic	764	55.31	43.15–63.76	3.38	141	54.26	45.82–62.68	3.59
Linolenic	726	6.18	0.89–11.08	2.57	140	5.43	0.89–11.01	2.92
Oil	714	20.28	12.50–28.10	5.87	116	18.35	13.75–23.57	2.01

[a] 2006 samples were separated from 2003 to 2006 samples because they were used as separate validation set

[b] The number of sample differs between instrument due of the presence of outliers (spectral and chemical)

[c] Fatty acid concentrations were expressed in relative concentrations while oil concentrations were expressed in absolute concentrations

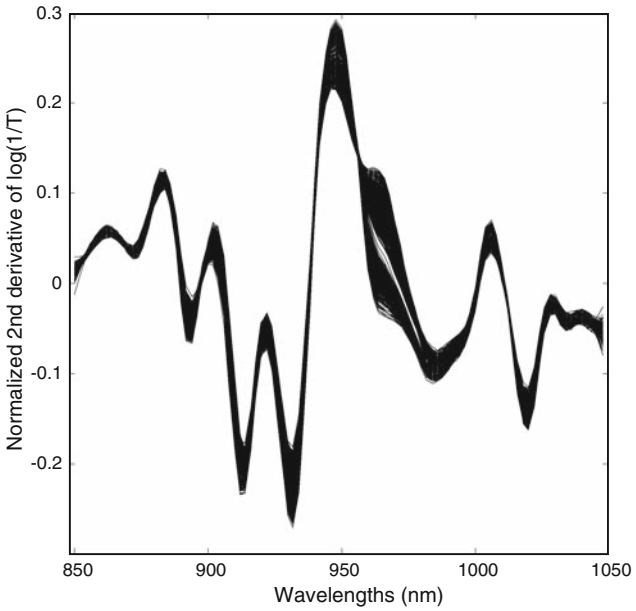

Fig. 1 Calibration set after second derivative and normalization. A scaling (mean zero and unit standard deviation) is necessary before implementing regression methods

Calibration Techniques

Partial least squares, a linear method, and artificial neural networks and least squares support vector machines, two non-linear regression techniques, were used to develop calibration models for the fatty acids.

Partial Least Squares Regression

Partial least squares regression extracts from the spectral data (**X**_matrix) the information that is related to the reference value of interest (**Y**_matrix). This extraction is performed by calculating principal components or latent variables that maximize the covariance between the **Y**_matrix and all possible linear functions of the **X**_matrix [28]. The choice of latent variable to include in the model is usually determined by minimizing the SECV and limiting overfitting. The use of PLS regression is rather simple and the number of samples required is not a limitation.

Artificial Neural Networks Regression

Artificial neural networks is a technique, primarily developed for classification, based on a network of individual interconnected neurons located on different layers (typically input, hidden and output layers). A weight and an activation function are associated to each neuron and it is the adaptation of these weights by back propagation that allows ANN to fit the input data. ANN presents three main drawbacks. The first is the amount of data needed (the number of observation must be larger than the number of weights to evaluate). The second difficulty lies in the number of parameters to tune (activation function, network structure, weight adaptation functions etc.) requiring mastery of the technique. Finally, the error plane of ANN can present local minima that do not represent the best fit of the training data. A solid validation strategy must be used to limit under and overfitting. The theory and applications of ANN to NIRS can be found in Williams and Norris [29].

Least Squares Support Vector Machines

Least squares support vector machines was also developed for classification purposes. LS-SVM has been developed to perform accurately on data presenting non-linear relationships with a limited number of observations. Similarly to support vector machines classification that looks for the maximum margin between clusters, LS-SVM tries to minimize the prediction error relative to an error rate determined by the user. The main advantage of LS-SVM is

that only two parameters need to be determined: the error rate and the parameter of the kernel function. The error plane presents only one minimum. However, its main drawback is the computation time; it is exponentially proportional to the size of the dataset and can take several hours to perform on a set of several hundred samples. Cogdill and Dardenne [30] provided a good overview of LS-SVM. Kecman [31] is a reference for theoretical aspects of both ANN and SVM.

Calibration and Validation Procedures

MATLAB R2007a (The MathWorks, Natick, MA) was used to support all calculations. The pretreatment of the data, the outlier detection and the development of PLS models were performed with the PLS_toolbox 4.0.2 (Eigenvector Research, Wenatchee, WA); ANN prediction models were created using the neural network toolbox v. 5.0.2 provided with MATLAB; and LS-SVM calibrations were developed with the LS-SVMlab toolbox v. 1.5 for MATLAB by Pelckmans et al. [32].

The ANN were created by first applying a principal component analysis to the pretreated data. The networks had one hidden layer with a tangent sigmoid transfer function and a linear activation function on the output layer. The number of neurons as well as the tuning of the parameters was based on the SEP calculated from independent validation sets (a description of the different validation sets is given later in the text). To limit overfitting, a randomly selected stop set representing 10% of the calibration set was used.

Parameters of LS-SVM were calculated using an optimization by exhaustive search on 25% of the calibration set. Locally optimized parameters were applied on the entire dataset and validated. This operation reduced the calculation time without deteriorating the SEP.

Validation Strategies

Three validation strategies were intended. Igne et al. [33] demonstrated the impact of the variability of next year samples on the calibration process. Moreover, no study on the measurement of fatty acids on whole soybean by NIRS attempted to validate a model on next year samples. Thus, for each instrument, each fatty acid and each regression technique, three calibration sets and three validation sets were created. The first scenario used samples from 2003 to 2005 where 80% were used for calibration and 20% for validation (Scenario 1). The second used all samples from 2003 to 2005 for calibration and models were validated on 2006 crop year samples (Scenario 2). The last scenario used 80% of the samples from 2003 to 2006 for calibration and the remaining 20% were used for validation (Scenario 3). These validation strategies were intended to evaluate the impact of including new samples in the calibration pool for parameters subject to an important variability from year to year due to breeding strategies.

Model Evaluation and Comparison Parameters

SEP was used to evaluate the precision of each model. SEP is the standard deviation of differences between the **Y**_matrix of validation and the prediction matrix $\hat{y}$. The models' fit was evaluated using the coefficient of determination (r^2) that represents the percentage of variability explained by the model.

Fatty Acid Calibration on Absolute and Relative Concentration

Two types of models were tested: models with fatty acids expressed in relative concentration (grams of fatty acids per 100 g of oil) and models with fatty acids expressed in absolute concentration (grams of fatty acid per 100 g of grain). The first situation predicted directly in percent of oil, the industry standard. The second used absolute concentrations during the calibration process, but absolute predictions were transformed to relative concentrations by multiplying them with the predicted oil content of the same sample.

Results and Discussion

Fatty Acids Expressed as Percentage of Oil

PLS, ANN, and LS-SVM validation results for the four instruments were averaged because the difference within instruments for the prediction of the same fatty acid was not significant (95% confidence interval). Table 2 presents validation results for each regression method, each fatty acid, and each validation scenario.

Validation Scenario 1 and 3

When comparing coefficients of determination of validation scenarios 1 and 3, where the variability of the validation set is more or less present in the calibration set, with those obtained by Kovalenko et al. [22] who performed the same type of validation strategies, we observed a good agreement for saturated and linoleic acids (Kovalenko et al. used spectra from 1991 to 2003; only the 2003 spectra were reused in this study). Predicted values of palmitic, stearic, and linolenic acids were in very good agreement with chemistry values while Kovalenko et al. reported lower r^2. The relationship between predicted and

Table 2 Validation statistics of the six fatty acids from calibrations developed with fatty acids expressed in relative concentration

Regression	Validation scenarios	Fatty acids											
		Saturated		Palmitic		Stearic		Oleic		Linoleic		Linolenic	
		r^2	SEP[d]	r^2	SEP	r^2	SEP	r^2	SEP	r^2	SEP	r^2	SEP
PLS[e]	Scenario 1[a]	0.97	0.72	0.97	0.64	0.85	0.30	0.59	1.62	0.77	1.51	0.95	0.64
	Scenario 2[b]	0.76	1.26	0.59	1.24	0.09	0.66	0.48	1.77	0.50	2.07	0.77	1.44
	Scenario 3[c]	0.95	0.85	0.93	0.85	0.66	0.42	0.60	1.64	0.72	1.67	0.91	0.90
ANN[f]	Scenario 1	0.98	0.57	0.98	0.49	0.86	0.29	0.56	1.70	0.79	1.45	0.95	0.52
	Scenario 2	0.70	1.18	0.66	1.08	0.08	0.69	0.35	2.00	0.46	2.17	0.77	1.41
	Scenario 3	0.97	0.69	0.96	0.62	0.68	0.40	0.59	1.64	0.70	1.71	0.83	0.66
LS-SVM[g]	Scenario 1	0.97	0.67	0.97	0.61	0.86	0.29	0.57	1.40	0.78	1.48	0.93	0.64
	Scenario 2	0.71	1.16	0.66	1.07	0.07	0.76	0.57	1.49	0.62	1.60	0.77	1.41
	Scenario 3	0.96	0.77	0.94	0.80	0.73	0.37	0.59	1.65	0.72	1.66	0.89	0.86

[a] 80% of samples from 2003 to 2005 constitute the calibration set, the 20% remaining form the validation set (e.g., 2006 samples not used)

[b] Samples from 2003 to 2005 constitute the calibration set, samples from 2006 the validation set

[c] 80% of samples from 2003 to 2006 constitute the calibration set, the 20% remaining form the validation set

[d] Standard error of prediction, expressed in % of oil

[e] Partial least squares

[f] Artificial neural networks

[g] Least squares support vector machines

chemistry values was the weakest for oleic acid and lower than what the literature reports. However when comparing SEPs, we observed a significant improvement over reports in the literature. For PLS, and validation scenario 3, SEPs were reduced by 61, 73, 57, 62, 57 and 48% for saturated, palmitic, stearic, oleic, linoleic, and linolenic acids respectively.

Validation Scenario 2

This validation strategy, representing conditions of commercial application of the models, revealed that the variability from year to year for fatty acids profiles was a major source of error. While saturated and linolenic acid models performed fairly well when predicting 2006 crop samples with PLS ($r^2 = 0.76$ and 0.77 respectively), palmitic, oleic, and linoleic acids calibrations yielded r^2 values barely usable for screening and the stearic acid model was not utilizable. In terms of precision, SEPs were globally doubled compared to the previous validation scenario but remained always under SEPs provided by Kovalenko et al. This situation is a further proof that the annual variability of composition is important, and more samples are needed in the calibration set to limit the increase of SEP. This also shows that the true test for a prediction model is its ability to predict totally independent samples. Cross-validation and validations on similar samples are only approximations of the model performances.

Comparison Between Regression Techniques

Figure 2 presents SEPs, averaged per instrument, from PLS, ANN, and LS-SVM models for each fatty acid and validation scenarios 2 and 3. The effect of the regression methods was parameter and validation strategy dependent. Table 3 summarizes the differences between validation strategies and regression methods. In eight cases out of twelve, LS-SVM gave the lowest SEPs but these results were significant for only three situations ($\alpha = 5\%$). ANN gave significantly lower SEPs in three cases and PLS in one case ($\alpha = 5\%$). It is interesting to note that PLS was the best technique for the validation of the stearic model by next year samples, a parameter that is hardly predictable by NIRS in the presented circumstances.

Fatty Acids Expressed in Absolute Concentration and Corrected to Relative Concentration by NIRS Oil Prediction

The use of absolute fatty acid concentration was suggested by Dr. Steven Wright, Pioneer Hi-Bred International Inc. [34] because of the nature of NIRS to determine content by counting molecules. Soybean oil calibrations were developed on the fatty acid calibration samples. Table 4 presents the different oil prediction model statistics. Predictions by these models were multiplied by predicted fatty acid concentrations on absolute scale to obtain fatty acids in grams per 100 g of oil. Validation statistics were averaged by

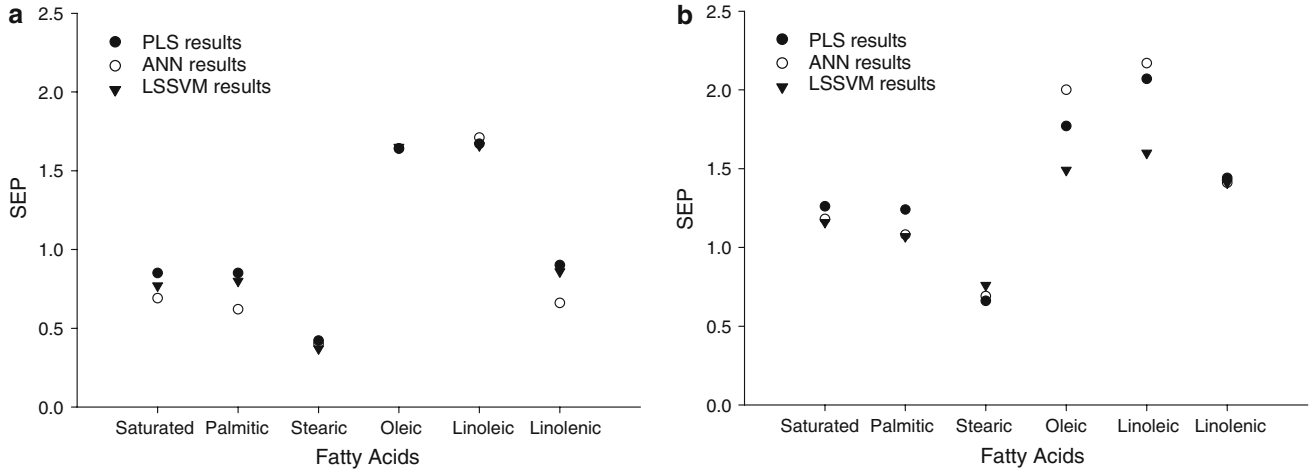

Fig. 2 Comparison of the precision for the different regression models [Partial least squares (PLS), Artificial neural networks (ANN), and Least squares support vector machines (LS-SVM)]. **a** Validation strategy where 2006 crop samples were predicted by a calibration set with samples from 2003 to 2005. **b** Validation strategy where the validation set was constituted by 20% of the samples from the original calibration set (samples from 2003 to 2006 crop years)

Table 3 Difference between regression methods for fatty acid prediction

Fatty acids	Validation strategy	SEPs[a]		
		Lower ⟷		Higher
Saturated	Scenario 2[b]	LS-SVM*	ANN*	PLS
	Scenario 3[c]	ANN**	LS-SVM*	PLS
Palmitic	Scenario 2[b]	LS-SVM	ANN	PLS
	Scenario 3[c]	ANN*	LS-SVM	PLS
Stearic	Scenario 2[b]	PLS*	ANN*	LS-SVM
	Scenario 3[c]	LS-SVM	ANN	PLS
Oleic	Scenario 2[b]	LS-SVM**	PLS*	ANN
	Scenario 3[c]	LS-SVM	ANN	PLS
Linoleic	Scenario 2[b]	LS-SVM*	PLS	ANN
	Scenario 3[c]	LS-SVM	PLS	ANN
Linolenic	Scenario 2[b]	LS-SVM	ANN	PLS
	Scenario 3[c]	ANN*	LS-SVM	PLS

PLS partial least squares, *ANN* artificial neural networks, *LS-SVM* least squares support vector machines

* Significant difference

** Significantly different on its own

[a] Standard Error of Prediction

[b] Samples from 2003 to 2005 constitute the calibration set, samples from 2006 the validation set

[c] 80% of samples from 2003 to 2006 constitute the calibration set, the 20% remaining form the validation set

Table 4 Statistics for total oil prediction models for each instrument and regression method

Instruments	PLS[a]		ANN[b]		LS-SVM[c]	
	r^2	SECV[d]	r^2	SECV	r^2	SECV
Infratec 1229	0.99	0.16	0.99	0.12	0.99	0.11
Infratec 1241	0.98	0.18	0.98	0.18	0.98	0.17
OmegAnalyzerG 106110	0.98	0.17	0.98	0.15	0.98	0.16
OmegAnalyzerG 106118	0.98	0.18	0.99	0.16	0.99	0.15

[a] Partial least squares

[b] Artificial neural networks

[c] Least squares support vector machines

[d] Standard error of cross validation

instrument for each validation strategy and each fatty acid because results among instruments were not significantly different ($\alpha = 5\%$). Figure 3 reports, for validation scenario 3 (80% calibration, 20% validation with all years included), SEPs in relative concentration and SEPs derived from absolute concentration (predictions in % of weight corrected to % of oil by multiplying by the oil content) as well as r^2 of fatty acid models developed on absolute concentration (r^2_a), r^2 of fatty acid models developed on relative concentration (r^2) and r^2 of models developed on relative concentration derived from predicted oil (r^2_r). Only LS-SVM results were reported since it was the overall best regression method and trends observed with PLS and ANN were similar.

For stearic, oleic, and linoleic acids, r^2_r was significantly lower than r^2_a; for palmitic and linolenic, coefficients of determination were not significantly different; and for saturated, r^2_r was significantly higher than r^2_a.

Also, when comparing r^2_a with r^2, there was significant improvement of the prediction models for oleic and linoleic acids when developing models on absolute value with PLS. However, when comparing r^2_r with r^2 and SEPs, models based on relative concentrations performed equivalently for saturated, palmitic and linolenic acids, worse for oleic

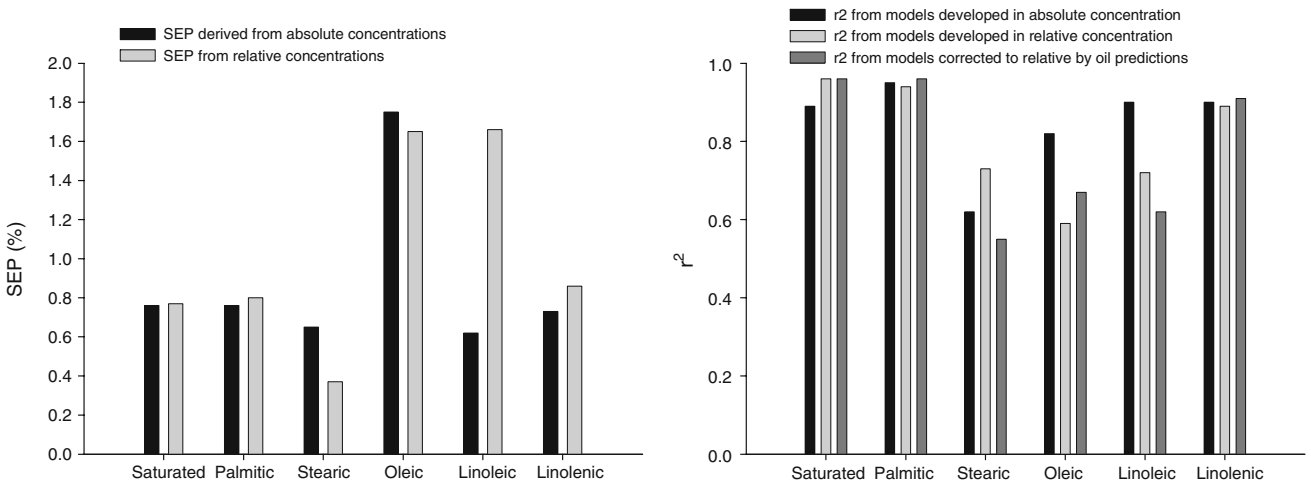

Fig. 3 Validation statistics from least squares support vector machines calibrations developed with fatty acids expressed in absolute and relative concentration. *SEP* Standard error of prediction

acid, and better for stearic acid. No changes were observed for linolenic acid.

While models developed on absolute concentrations might have had a better precision, since fatty acid values were converted to relative units, SEPs were often equivalent, even higher for some elements. This situation is probably caused by the compounded error in the oil and absolute fatty acid models; errors of oil models, as low as they were, were multiplied by the error of fatty acid predictions on absolute concentration and as a consequence, increased the total error of the model given in relative concentration. Changing market practice from relative to absolute concentrations would improve NIRS results. NIRS calibration models developed on absolute concentration using oil and fatty acid contents measured by reference methods would limit the impact of the oil prediction error in the final result.

Local Regression for Linolenic Acid

Among fatty acids described here, linolenic acid is particularly of interest for breeders. Figure 4 presents the calibration results of soybean linolenic acid. This shows clearly that two clusters are mostly represented: high and low linolenic concentrations. While the calibration curve for linolenic acid concentration between 6 and 10% looks good, difficulties are apparent for concentrations lower than 3%. At this point, NIRS can be used to screen high and low linolenic samples. To investigate the possibility to predict more precisely between 1 and 3%, a local PLS model on Infratec 1229 was developed with these "low linolenic" values. Figure 5 presents the calibration and the validation curves for linolenic acid with values up to 3% and the validation of the model with scenario 3 [80% of

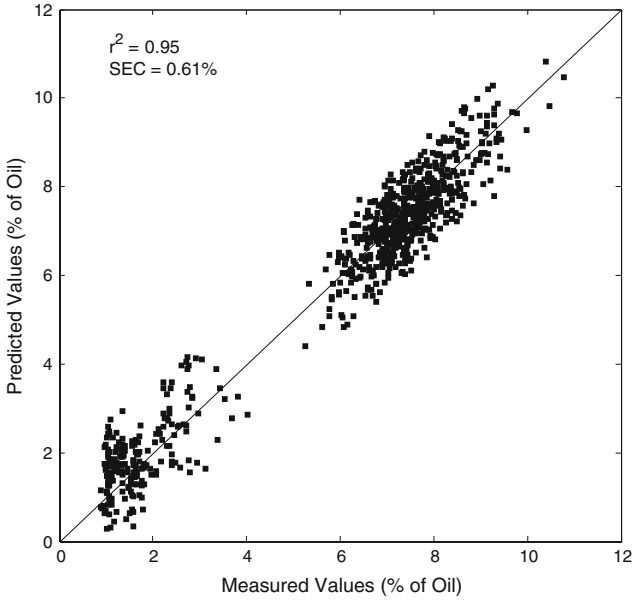

Fig. 4 Regression curve for linolenic acid using partial least squares regression method (Infratec 1229 spectral data). *SEC* Standard error of calibration

samples from 2003 to 2006 in calibration (208 samples)], and the 20% remaining in the validation set (51 samples). The r^2 is 0.52 and the SEP is 0.42%. The model precision is not as good as those obtained when predicting the entire range. However, the values measured by chemistry between 0.9 and 1.2% were not well predicted. This situation shows the difficulty that NIRS faces when predicting low linolenic acid concentrations. When using the validation scenario 2 (2003–2005 in calibration, 2006 samples in validation), we obtained completely different results with a r^2 of 0 and a SEP of 0.78% (Fig. 6). This would signify that more samples would be necessary to improve the

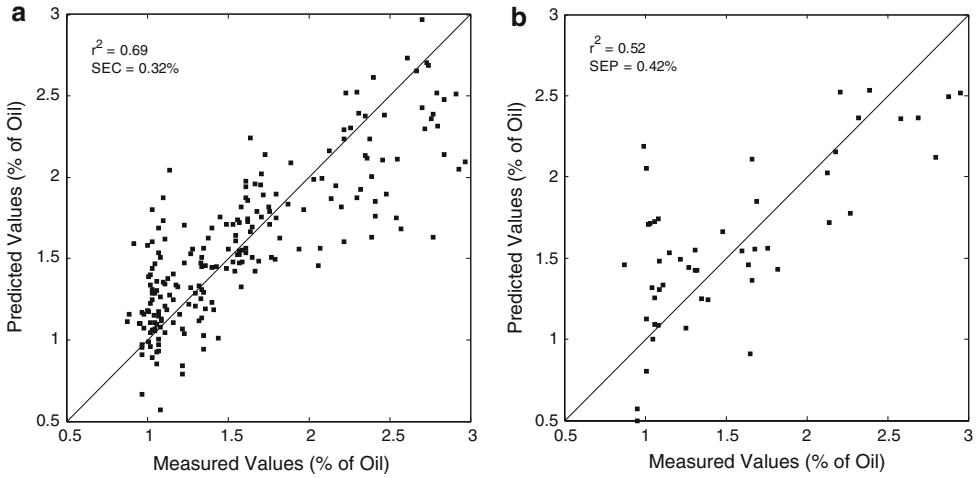

Fig. 5 Regression (**a**) and validation (**b**) curves for low linolenic acid samples (up to 3%). The validation was performed with 20% of the samples of the calibration set. *SEC* Standard error of calibration, *SEP* standard error of prediction

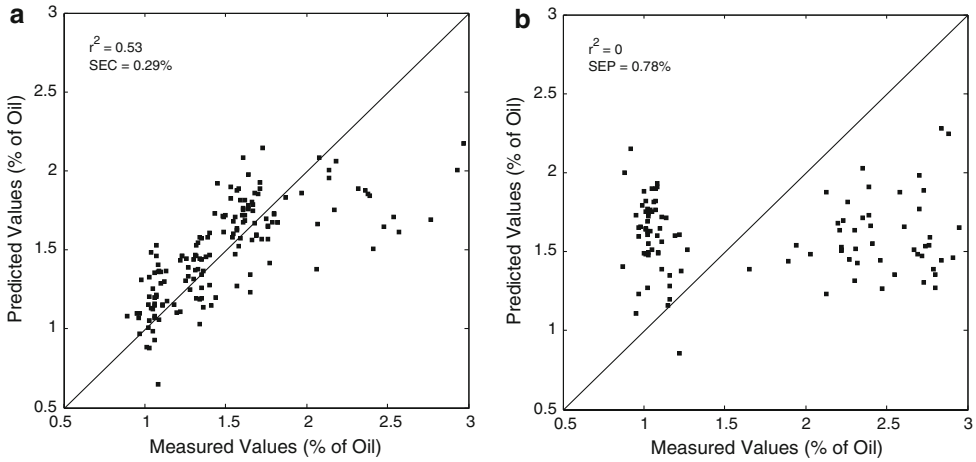

Fig. 6 Regression (**a**) and validation (**b**) curves for low linolenic acid samples (up to 3%). The validation was performed with next year (2006) samples. *SEC* Standard error of calibration, *SEP* standard error of prediction

calibration at low linolenic acid concentrations, if possible at all on whole grain analysis.

Conclusions

Near infrared spectroscopy is becoming a tool of choice for the determination of whole soybean fatty acid profiles. This study presented a lower error of prediction than previous studies. Non-linear regression methods appeared as a necessary improvement for the development of the technology. The variability of the sample included in calibration and validation was proven to impact prediction model statistics. Models should undergo a validation with external samples since cross validation methods give only an approximation of the real performances. Instrument software will need upgrades because no instrument supports all nonlinear methods in their operating routines.

Comparison between models developed on relative and absolute fatty acid concentrations showed that even though models were easier to develop using absolute concentrations, the error generated by the prediction of oil used to correct absolute predictions to relative contents increased the final error. These results are partially in agreement with Wright et al. (2003) showing that the non-linear relationship between fatty acid content and absorbance could be corrected by using absolute concentrations. However, the correction to relative concentrations by the NIRS-predicted oil content degraded the calibration precision. Present results indicate that absolute concentrations could be beneficial for the fast screening and quality control of soybean samples.

References

1. Jakobsen MU, Overvad K, Dyerberg J, Heitmann BL (2004) Dietary fat and risk of coronary heart disease: possible effect modification by gender and age. Am J Epidemiol 160:141–149
2. Posner BM, Cobb JL, Belanger AJ, Cupples LA, D'Agostino RB, Stokes J (1991) Dietary lipid predictors of coronary heart disease in men. The Framingham Study. Arch Intern Med 151:1181–1187

3. Hu FB, Stampfer MJ, Manson JE, Rimm E, Colditz GA, Rosner BA, Hennekens CH, Willett WC (1997) Dietary fat intake and the risk of coronary heart disease in women. N Engl J Med 337:1491–1499
4. Xu J, Eilat-Adar S, Loria C, Goldbourt U, Howard BV, Fabsitz RR, Zephier EM, Mattil C, Lee ET (2006) Dietary fat intake and risk of coronary heart disease: the strong heart study. Am J Clin Nutr 84:894–902
5. Ziedén B, Kaminskas A, Kristenson M, Olsson AG, Kucinskiene Z (2002) Long chain polyunsaturated fatty acids may account for higher low-density lipoprotein oxidation susceptibility in Lithuanian compared to Swedish men. Scand J Clin Lab Invest 62:307–314
6. Ip C (1997) Review of the effects of trans fatty acids, oleic acid, n-3 polyunsaturated fatty acids, and conjugated linoleic acid on mammary carcinogenesis in animals. Am J Clin Nutr 66(suppl):1523S–1529S
7. Wijendran V, Hayes KC (2004) Dietary n-6 and n-3 fatty acid balance and cardiovascular health. Annu Rev Nutr 24:597–615
8. List GR, Mounts F, Orthoefer F, Neff WE (1996) Potential margarine oils from genetically modified soybeans. J Am Oil Chem Soc 73:729–732
9. Fehr WR, Welke GA, Hammond E, Cianzio S (2001) Inheritance of elevated palmitic acid content in soybean seed oil. Crop Sci 32:1522–1524
10. Weaver CM, Mason AC, and Hamaker BR (2000) Food uses. In: Designing crops for added value, ASA, CSSA, SSSA, Madison, WI, USA, pp 21–55
11. Stender S, Dyerberg J (2004) Influence of trans fatty acids on health. Ann Nutr Metab 48:61–66
12. Durham D (2003) The United Soybean Board's better bean initiative: Building United States soybean competitiveness from the inside out. AgBioForum, 6:23–26. http://www.agbioforum.org (accessed June 2006)
13. Rippke GR, Hardy CL, Hurburgh CR Jr (1996) Calibration and field standardization of Tecator Infratec analyzers for corn and soybeans. In Near Infrared Spectroscopy: The Future Waves, NIR Publications, Chichester, UK, pp122–131
14. Delwiche SR (2004) Analysis of Small Grain Crops. In Near-Infrared Spectroscopy in Agriculture, ASA, CSSA, SSSA, Madison, Wisconsin, USA, pp 269–320
15. Kovalenko IV, Rippke GR, Hurburgh CR (2006) Determination of Amino Acid Composition of Soybeans (gram lysine max) by Near-Infrared Spectroscopy. J Agric Food Chem 54:3485–3491
16. Igne B, Gibson LR, Rippke GR, Schwarte A, Hurburgh CR Jr (2007) Triticale moisture and protein prediction by near infrared spectroscopy. Cereal Chem 84:328–330
17. Velasco L, Becker HC (1998) Estimating the fatty acid composition of the oil in intact-seed rapeseed (*Brassica napus* L.) by near-infrared reflectance spectroscopy. Euphytica 101:221–230
18. Pérez-Vich B, Velasco L, Fernández-Martínez JM (1998) Determination of seed oil content and fatty acid composition in sunflower through the analysis of intact seeds, husked seeds, meal and oil by near-infrared reflectance spectroscopy. J Am Oil Chem Soc 75:547–555
19. Tillman BL, Gorbet DW, Person G (2006) Predicting oleic and linoleic acid content of single peanut seeds using near-infrared reflectance spectroscopy. Crop Sci 46:2121–2126
20. Pazdernik DL, Killam AS, Orf JH (1997) Analysis of amino and fatty acid composition in soybean seed, using near infrared reflectance spectroscopy. Agron J 89:679–685
21. Nimaiyar S, Paulsen MR, Nelson RL (2004) Rapid analysis of fatty acids in soybeans using FT-NIR. Paper Number 046118, ASAE/CSAE Annual International Meeting, Ottawa, ON, Canada
22. Kovalenko IV, Rippke GR, Hurburgh CR (2006) Measurement of soybean fatty acids by near-infrared spectroscopy: linear and nonlinear calibration methods. J Am Oil Chem Soc 83:421–427
23. US National Type Evaluation Program, Certificate Number 06-081A1, http://www.ncwm.net/ntep/certificate/06-081A1.pdf
24. US National Type Evaluation Program, Certificate Number 95-063A11, http://www.ncwm.net/ntep/certificate/95-063A11.pdf
25. US National Type Evaluation Program, Certificate Number 01-063A8, http://www.ncwm.net/ntep/certificate/01-063A8.pdf
26. Hammond EG (1991) Organization of rapid analysis of lipids in many individual plants. In: Modern methods of plant analysis, new series, vol 12, Essential oils and waxes. Springer-Verlag, New York, pp 321–330
27. Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1639
28. Naes T, Isakson T, Fearn T, Davies T (2002) Multivariate calibration and classification. NIR Publications, Chichester
29. Williams P, Norris K (2001) Near infrared technology in the agricultural and food industry, 2nd edn. AACC Inc., St Paul
30. Cogdill RP, Dardenne P (2004) Least-squares support vector machines for chemometrics: an introduction and evaluation. J Near Infrared Spec 12:93–100
31. Kecman V (2001) Learning and soft computing. Cambridge, London
32. Pelckmans K, Suykens JAK, Van Gestel T, De Brabanter J, Lukas L, Hamers B, De Moor B, Vandewalle J (2003) LS-SVMlab toolbox. University of Leuven, Belgium
33. Igne B, Gibson LR, Rippke GR, Hurburgh CR Jr (2007) Influence of the yearly variability of agricultural products on the calibration process. Cereal Chem 84:576–581
34. Wright S, Hagen L (2003) Oleic acid content in ground corn by NIR spectroscopy with an indirect calibration method. JAOCS 80:1163–1167