LWT - Food Science and Technology 79 (2017) 126–134

Contents lists available at ScienceDirect

LWT - Food Science and Technology

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

Estimation of composition of quinoa (*Chenopodium quinoa* Willd.) grains by Near-Infrared Transmission spectroscopy

Christian Encina-Zelada [a, b, c], Vasco Cadavez [a], Jorge Pereda [b], Luz Gómez-Pando [d], Bettit Salvá-Ruíz [b], José A. Teixeira [c], Martha Ibañez [d], Kristian H. Liland [e], Ursula Gonzales-Barron [a, *]

[a] *CIMO Mountain Research Centre, School of Agriculture, Polytechnic Institute of Braganza, Portugal*
[b] *Department of Food Technology, Faculty of Food Industries, National Agricultural University La Molina, Lima, Peru*
[c] *Department of Biological Engineering, School of Engineering, University of Minho, Portugal*
[d] *Cereals and Andean Crops Programme, Faculty of Agronomy, National Agricultural University La Molina, Lima, Peru*
[e] *Nofima AS — Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, N-1430, Ås, Norway*

ARTICLE INFO

Article history:
Received 23 November 2016
Received in revised form
9 January 2017
Accepted 10 January 2017
Available online 11 January 2017

Keywords:
Canonical
Partial least squares
Chemometrics
Scatter correction
Savitzky-Golay

ABSTRACT

The aim of this study was to develop robust chemometric models for the routine determination of dietary constituents of quinoa (*Chenopodium quinoa* Willd.) using Near-Infrared Transmission (NIT) spectroscopy. Spectra of quinoa grains of 77 cultivars were acquired while dietary constituents were determined by reference methods. Spectra were subjected to multiplicative scatter correction (MSC) or extended multiplicative signal correction (EMSC), and were (or not) treated by Savitzky-Golay (SG) filters. Latent variables were extracted by partial least squares regression (PLSR) or canonical powered partial least squares (CPPLS) algorithms, and the accuracy and predictability of all modelling strategies were compared. Smoothing the spectra improved the accuracy of the models for fat (root mean square error of cross-validation, RMSECV: 0.319–0.327%), ashes (RMSECV: 0.224–0.230%), and particularly for protein (RMSECV: 0.518–0.564%) and carbohydrates (RMSECV: 0.542–0.559%), while enhancing the prediction performance, particularly, for fat (root mean square error of prediction, RMSEP: 0.248–0.335%) and ashes (RMSEP: 0.137–0.191%). Although the highest predictability was achieved for ashes (SG-filtered EMSC/PLSR: bootstrapped 90% confidence interval for RMSEP: [0.376–0.512]) and carbohydrates (SG-filtered MSC/CPPLS: 90% CI RMSEP: [0.651–0.901]), precision was acceptable for protein (SG-filtered MSC/CPPLS: 90% CI RMSEP: [0.650–0.852]), fat (SG-filtered EMSC/CPPLS: 90% CI RMSEP: [0.478–0.654]) and moisture (non-filtered EMSC/PLSR: 90% CI RMSEP: [0.658–0.833]).

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Quinoa (*Chenopodium quinoa* Willd.) is a pseudocereal originating from the surroundings of the Titicaca Lake (Peru and Bolivia), which has been cultivated for centuries in the Andean countries. Quinoa is known as a pseudo-cereal because its seeds are used as cereal grains; although its nutritional quality is superior to that of the common cereals (Jancurová, Minarovicová, & Dandar, 2009; Vega-Gálvez et al., 2010).

Near infrared transmission (NIT) spectroscopy can presently provide rapid and accurate analysis of starch, moisture, protein, and oil contents in whole kernel cereals (Büchman, Josefsson, & Cowe, 2001; Miralbés, 2004; and; Pojić, Mastilović, Pestorić, & Radusin, 2008). However, when analysing intact samples by diffuse reflectance or transmittance spectroscopy, uncontrolled variations in light scattering are often a dominating artifact that complicates subsequent chemometric modelling (Panero, Panero, Panero, & Silva, 2013). This undesired scattering variation is due to uncontrolled physical variations of the samples, such as particle size and shape, sample packing, surface and orientation of the particles (Cantor, Hoag, Ellison, Khan, & Lyon, 2011). In order to minimise the multiplicative interference of scatter and particle size for the construction of robust models, NIT spectra are subjected to processing techniques for signal correction (i.e., multiplicative scatter correction and extended multiplicative signal correction) and noise

[*] Corresponding author. School of Agriculture, Polytechnic Institute of Braganza, Campus de Santa Apolónia, Apartado 1172, 5301-854, Portugal.
E-mail address: ubarron@ipb.pt (U. Gonzales-Barron).

http://dx.doi.org/10.1016/j.lwt.2017.01.026
0023-6438/© 2017 Elsevier Ltd. All rights reserved.

removal (i.e., Savitzky-Golay derivatives).

Processed spectroscopy data matrices are then related with physicochemical data using multivariate calibration methods (Ferreira, Pallone, & Poppi, 2015). Partial least squares regression (PLSR) is currently considered as one of the most robust multivariate regression techniques as it is associated with prediction errors that are lower than those of the principal component analysis (Moghimi, Aghkhani, Sazgarnia, & Sarmad, 2010; Wold, Martens, & Wold, 1983). Recently, a generalisation of PLSR has been proposed that incorporates discrete and continuous responses, additional measurements, and individual weighting of observations. The technique is known as Canonical Powered Partial Least Squares (CPPLS) because the optimal latent variables are found by combining PLS methodology and canonical correlation analysis (Indahl, Liland, & Næs, 2009; Mevik, Wehrens, & Liland, 2015). Thus, the objective of this study was three-fold: (i) to assess the feasibility of accurately quantifying dietary constituents of quinoa (moisture, protein, fat, ashes and carbohydrates) whole grains by NIT spectroscopy; (ii) to compare the robustness and prediction capability of the PLSR and CPPLS multivariate models after scatter correction of the spectra; and (iii) to assess to what extent smoothing filters applied to scatter-corrected spectra can further improve the performance of the PLSR and CPPLS algorithms.

2. Methodology

2.1. Samples and proximate composition analysis

The samples utilised in this study were quinoa (*Chenopodium quinoa* Willd.) whole grains of orange, beige, black and yellow colour, corresponding to 77 different cultivars. They were all harvested in Peru at the National Agricultural University La Molina (Lima) and the Regional Development Centre – Highland (Junin), between 2010 and 2012. Moisture, protein, fat and ashes contents were determined in triplicate using the reference methods 925.10, 920.87 (conversion factor of 6.25), 923.05 and 923.03, respectively, as described by the Association of Official Agricultural Chemists (AOAC, 2000). Total carbohydrate content was calculated by difference as: 100 - (weight in grams [protein + fat + water + ashes] in 100 g of quinoa). Proteins, fat, ashes and carbohydrate contents were then converted into dry basis (db).

2.2. Near-infrared transmission (NIT) spectra acquisition

NIT spectra were acquired by placing the whole grains directly in an Infratec 1241 grain analyser (Module Foss Tecator, Denmark), using 60-mm quartz cuvettes, and scanning the region 850–1048 nm (wavenumber range of 11,765–9524 cm^{-1}). The spectra were recorded at scanning step intervals of 2 nm to give 100 data points per sample. A total of 10 frequency scans were performed per sample, and carefully assessed for consistency. Raw spectral data (i.e., a vector of 100 data points per sample) were linked to the chemical analyses data on a spreadsheet. To correct for the non-linearity in the measure of transmittance (T), T was transformed into absorbance (A) by taking the base 10 logarithm of the reciprocal of the transmittance values ($A = \log 1/T$).

2.3. NIT spectral pre-processing

To minimise the multiplicative effects of light scattering, spectra were subjected to multiplicative scatter correction (MSC) or extended multiplicative signal correction (EMSC). MSC is a transformation method used to compensate for additive and multiplicative effects in spectral data (Maleki, Mouazen, Ramon, & De Baerdemaeker, 2007). Both EMSC and MSC attempts to separate physical light scattering effects from chemical (vibrational) light absorbance, yet EMSC is a modification of the standard MSC which adds polynomials to the correction model in addition to the constant baseline effect and reference scaling of MSC (Martens & Stark, 1991; Panero et al., 2013). The basic EMSC with polynomials of degree 2 was applied. For each of the dietary constituents analysed, PLSR and CPPLS multivariate models were then fitted to the MSC- or EMSC- pre-processed spectra; thereby producing four treatments (MSC/PLSR, EMSC/PLSR, MSC/CPPLS and EMSC/CPPLS) which were compared in terms of predictability.

In addition, Savitzky-Golay (SG) derivative filters (Savitzky & Golay, 1964) were applied after correcting spectra for scattering (MSC or EMSC) to assess whether the predictive performance of the PLSR and CPPLS models could be further enhanced. SG smoothing performs a piece-wise polynomial fitting with specified polynomial degree (p), window length (w), and derivative order (m) to the spectrum. Thus, SG filters produced by all possible combinations of $m = \{1, 2\}$, $p = \{2, 3, 4\}$ and $w = \{3, 5, 7, 9, 11\}$ were applied to each of the MSC and EMSC scatter-corrected spectra.

2.4. Chemometric multivariate data analysis

The extraction of information from quinoa grain's pre-processed spectra to estimate moisture, protein, fat, ashes and carbohydrates contents was performed by the PLSR and CPPLS chemometric algorithms. For the CPPLS models estimating moisture content, the additional variables were protein, fat, ashes and quinoa cultivar. For the estimation of protein by CPPLS, the additional variables were moisture, fat, ashes and cultivar; whereas for the estimation of fat, the additional variables were moisture, protein and ashes. The additional variables for ashes content CPPLS models were moisture, fat and quinoa cultivar, while those for carbohydrates content were moisture, ashes and fat. Selection of the additional variables for each dietary constituent's CPPLS model was carried out by trial and error.

As a first step, the full data set was divided into a subset for calibration (~80% data, 62 samples) and the remaining ~20% (15 samples) for prediction or validation, by means of random split stratified by cultivar. PLSR and CPPLS were fitted separately to MSC and EMSC scatter-corrected spectra with and without SG filters. The performance of the different models (a model is defined as a combination of a pre-processing filter and a chemometric multivariate algorithm) was determined by *cross-validation* as an *internal calibration* method using the calibration data set. In our case, the leave-one-out (LOO) method was used. Briefly, in the LOO method, each sample is removed one at a time from the calibration set, a new calibration performed and a prediction score calculated for the sample removed. This procedure is repeated until every sample has been left out once. The performance of the model was assessed by the root mean square error of cross-validation (RMSECV), which is deemed as the best single estimate of the prediction capability of the model (González-Martín, Moncada, Fischer, & Escuredo, 2014; Mevik & Wehrens, 2007). Then, the optimal number of components of a model was selected at the first RMSECV local minimum, rather than the absolute minimum (to avoid overfitting). For such a number of components, the root mean square error of calibration (RMSEC) was computed. In addition, the coefficients of correlation between reference values and values fitted by cross-validation (R_{CV}) and the calibration model (R_C) were computed.

Following completion of the calibration, models were validated using the prediction data set. Model performance was evaluated by obtaining the root mean square error of prediction (RMSEP) and the coefficient of correlation (R_P) between reference values and those predicted by the model. For each of the four treatments (i.e., MSC/PLSR, EMSC/PLSR, MSC/CPPLS and EMSC/CPPLS), the SG filters leading to the highest accuracy were identified. To assess the best model(s) for each dietary constituent, the model had to present not only a low RMSE but also a high R. The entire NIT spectra analysis was conducted using the “pls” (Mevik et al., 2015), “emsc” (Liland, 2016) and the “prospectr” (Stevens & Ramirez-Lopez, 2013) packages implemented in the R software version 3.2.5 (R Core Team, 2016).

3. Results and discussion

3.1. Proximate composition analysis of quinoa

The values reported in this study for fat (5.35–7.78% db) and ashes (2.51–4.11% db; Table 1) were comparable to those reported by Repo-Carrasco-Valencia, Hellström, Pihlava, and Mattila (2010) for six ecotypes of similar Peruvian quinoa (fat: 4.36–7.59% db, and ashes: 2.57–3.44% db). However, they found considerably higher protein content (12.55–16.08% db) and lower carbohydrates content (67.13–77.02% db) than those found in this report (8.33–11.38% db; and 78.48–82.89% db, respectively). Analysing quinoa samples from Peru, Bolivia and Brazil, Ferreira et al. (2015) encountered substantially higher fat (6.19–15.52% db) and ashes (3.07–9.15% db) contents than those of our study. The variation in ashes are influenced by the dependence of the mineral content on type of soil and fertiliser application. Moisture is the compound most variable among published studies (from 8.26 to 11.51% in Repo-Carrasco-Valencia et al. (2010) up to 25.66–33.16% in Ferreira et al. (2015)) because it depends upon drying and storage of seeds. The standard deviations suggest that sufficient variation in the dietary compounds existed among the quinoa cultivars to develop chemometric models.

3.2. Pre-processing methods for signal correction and smoothing of quinoa's NIT spectra

The first step of signal pre-treatment is crucial as redundant information should be removed from the spectra. With corrected spectra, the repeatability and reproducibility of the chemometric multivariate model can be increased (Stevens & Ramirez-Lopez, 2013). In the first instance, the transmittance spectra of the quinoa grains without any processing pointed to the occurrence of multiplicative scaling effects (Fig. 1, top left), which were still present when spectra were transformed into absorbance (Fig. 1, top right). Such transformation is needed to move signal processing to a domain where Beer-Lambert's law applies and additive effects of compounds are linear. Light scattering, one of the main causes of multiplicative scale effects (i.e., scale differences) in spectral data, was corrected by both methods, MSC (Fig. 1, bottom left) and EMSC (Fig. 1, bottom right), although the application of EMSC yielded a better signal correction. Whereas MSC was developed to remove both scaling effects (a multiplicative factor) and baseline shift effects (an additive factor), EMSC was designed to allow the separation of multiplicative physical effects (path length, light scattering, etc.) from additive chemical effects (absorbance of analytes and interferants) and additive physical effects (temperature shifts, baseline variations, etc.) (Panero et al., 2013). Hence, additive effects, chemical and/or physical, must have been also present in the raw spectra.

In general, when SG first (SG1) and second (SG2) derivative filters were applied to either the MSC- or the EMSC-corrected spectra, the peaks below and above the baseline were emphasised. It was not unexpected that EMSC + SG pre-processing (Fig. 2, bottom) produced cleaner signals than MSC + SG pre-processing (Fig. 2, top), as EMSC yielded a better correction for light scattering and additive effects than MSC. However, whether the application of SG1 or SG2 pre-processing smoothing filter produces better signals should be determined by the resulting predictive capacity of the chemometric models.

Table 1
Summary statistics of the major dietary compounds of quinoa samples in % dry basis, except for moisture (% wet basis).

Compound	Minimum	Maximum	Mean	St. deviation
Moisture	9.17	13.41	10.55	0.86
Protein	8.33	11.38	9.88	0.77
Fat	5.35	7.78	6.54	0.42
Ashes	2.51	4.11	3.13	0.40
Carbohydrates	78.48	82.89	80.45	0.98

3.3. Comparisons between scatter correction methods and multivariate algorithms

For moisture, protein and ashes contents, regardless of the chemometric algorithm used (i.e., PLSR or CPPLS), the application of EMSC to the spectra produced lower errors (i.e., RMSECV) by up to ~4.8% in the case of protein, than those produced by MSC treatments (Table 2). Comparing EMSC and MSC performance, Panero et al. (2013) similarly found lower RMSEC and RMSEP values when applying the former scatter correction method on marzipan spectra for NIR determination of moisture. Correspondingly, for moisture, protein and ashes contents, correcting the signal scatter by EMSC led to higher R_{CV} values (range of 0.572–0.769) than those produced by the simpler MSC (0.564–0.742; Table 2). Considering that the models fitted to EMSC-processed spectra consistently led to fewer optimal components (3–7) than those fitted to MSC-processed spectra (4–8), it can be stated that EMSC, with their resulting lower cross-validation errors and higher cross-validation correlation coefficients, had a tendency to produce more robust models than MSC for the NIT determination of moisture, protein and ashes. Nevertheless, in the cases of fat and carbohydrates, irrespective of the algorithm used for model calibration, the behaviour was the opposite; this is, MSC-treated spectra yielded more robust chemometric models – as implied by their lower RMSECV and higher R_{CV} – than the EMSC-treated spectra did, although with at most one more component (Table 2). For fat and carbohydrates, EMSC may have overfitted the baseline such that chemical information was discarded along with the scatter correction.

The multivariate regression methods also affected the accuracy of prediction for the models. In the analyses of all dietary components, the CPPLS algorithm led invariably to a selection of fewer optimal components (3–5) than PLSR (6–8). This was an anticipated outcome since CPPLS was developed as a compression method for the extraction of more predictive information in the first few components than ordinary PLSR (Indahl et al., 2009). For this reason, within each dietary constituent, the models with the combination CPPLS/EMSC yielded the lowest optimal number of components (3–4) while the combination PLSR/MSC yielded the highest optimal number of components (7–8). For instance, for the protein constituent, the 8 optimal latent variables in the combination PLSR/MSC was brought down to 3 in the combination CPPLS/EMSC. In all dietary constituents – except fat – there was a clear effect of the multivariate regression on the RMSEC and RMSEP values, being the CPPLS algorithm associated to higher errors (Table 2).

With the exception of carbohydrates, when the quinoa grains' spectra were MSC scatter-corrected, the use of the PLSR or CPPLS algorithm produced very similar cross-validation errors (RMSECV) for the estimation of moisture (0.575; 0.579%), protein (0.614; 0.613%), fat (0.326; 0.325%) and ashes (0.231; 0.233%). However, the

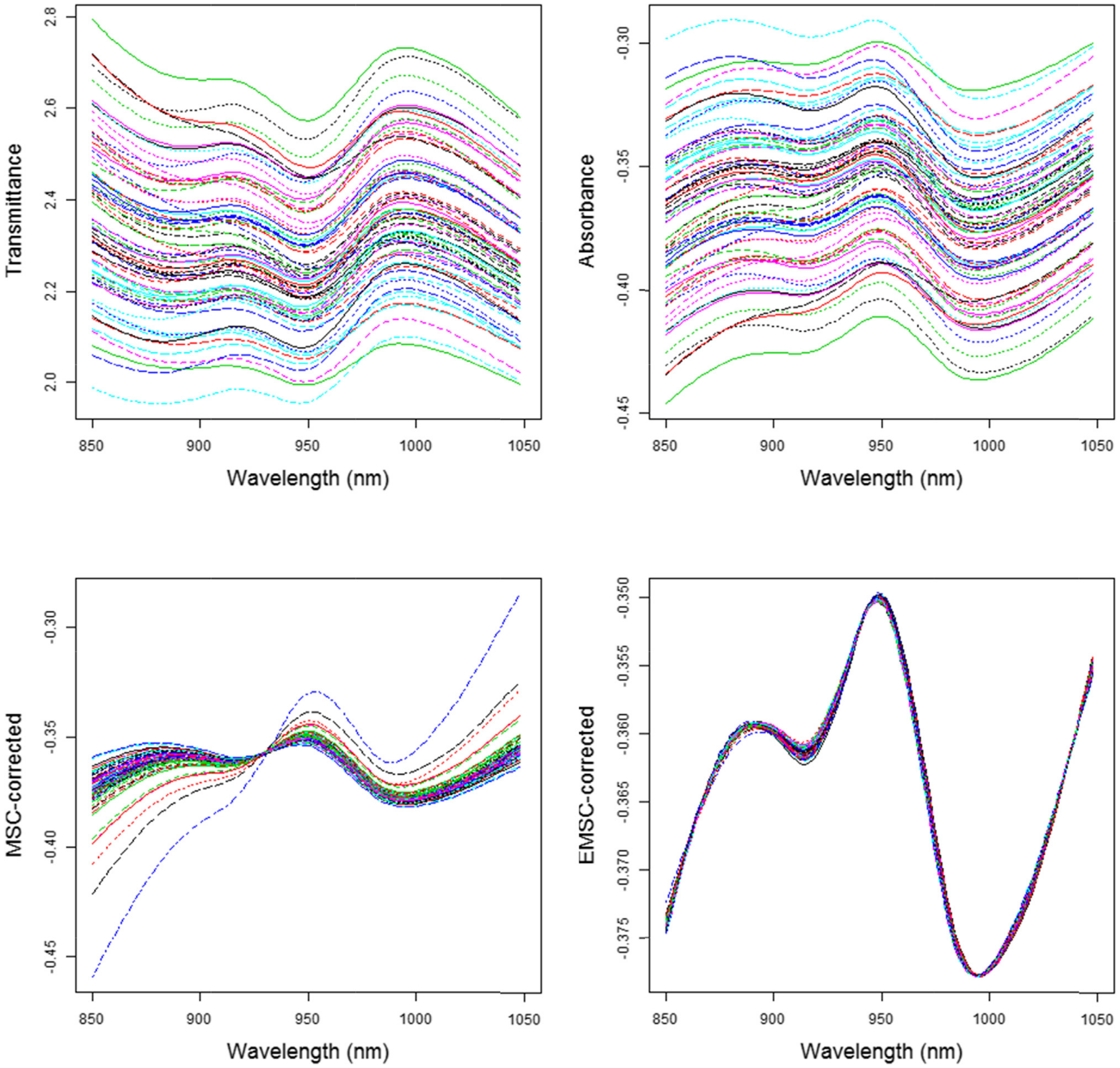

Fig. 1. Untransformed or raw near-infrared transmittance spectra of quinoa whole grains (top left), spectra transformed into absorbance (top right), and absorbance spectra corrected for scattering applying multiplicative scatter correction (MSC; bottom left) or extended multiplicative signal correction (EMSC; bottom right).

effect of the regression algorithm on RMSECV values became more noticeable when spectra were pre-processed by EMSC for the chemometric models determining moisture (RMSECV: 0.566; 0.578%) and carbohydrates (0.620; 0.638%). When applied to EMSC-treated spectra, the PLSR algorithm produced more accurate models – lower RMSECV in all dietary constituents – than those produced by CPPLS. Even for moisture, protein and ashes, the PLSR/EMSC treatment yielded the highest R_{CV} and R_C values among the four treatments. This may arise from the higher optimal number of components consistently picked by the PLSR algorithm (Table 2).

Earlier, Ferreira et al. (2015) proposed a series of chemometric models to estimate the proximate composition of quinoa from Fourier transform near-infrared (FTIR) spectra. In order to contrast the accuracy of our models with their FTIR models, the coefficient of variation (CV = RMSECV/mean) was calculated as a common metric for comparison since it is a dimensionless number less sensitive to difference in means. The chemometric models presented in this study were more accurate than those obtained in Ferreira et al. (2015), as indicated by the considerably lower CV of our models for moisture (5.3–5.5% as opposed to 5.9%), Ferreira et al. (2015) protein (5.8–6.2% as opposed to 14.9%), fat (4.9–5.2% as opposed to 11.7%), carbohydrates (0.73–0.79% as opposed to 7.0%) and ashes (7.0–7.4% as opposed to 15.5%). Similarly, the external validation CV (RMSEP/mean) obtained from our models for protein (5.5–6.4%) and fat (5.6–4.1%) were far lower than those reported by González-Martín et al. (2014) (10.4% and 8.3%, respectively). Nonetheless, when contrasting the estimates of correlation between the reference and the spectral methods, the R_{CV} (0.56–0.77) and R_C (0.51–0.83; Table 2) found in our models were, as a whole, lower than those reported by both González-Martín et al. (2014) (R_{CV}: 0.89–0.96) and Ferreira et al. (2015) (R_C: 0.86–0.91). The lower correlation coefficients encountered in this

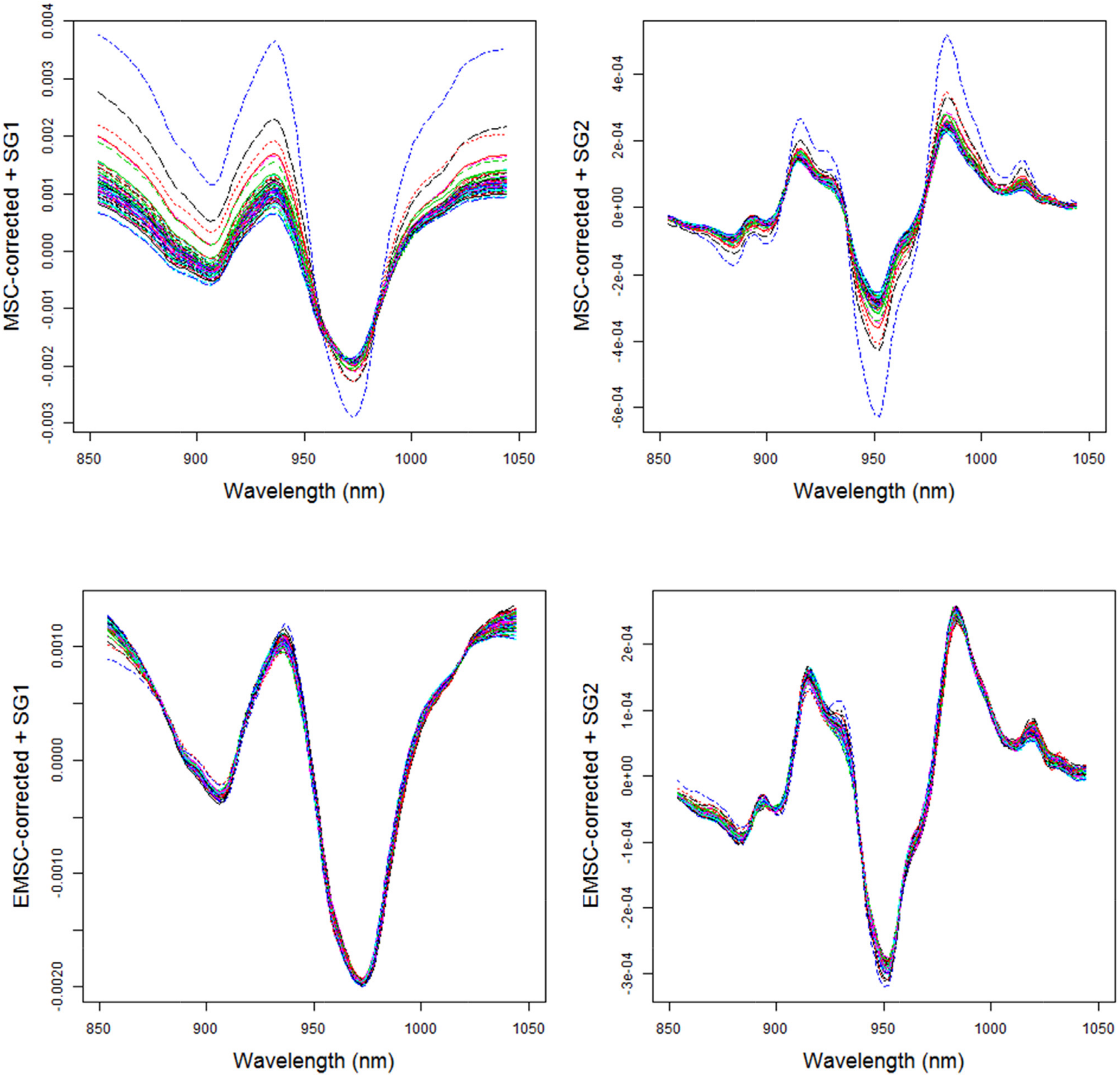

Fig. 2. Effects of applying Savitzky-Golay first- (SG1; left) and second-derivative (SG2; right) with polynomial degree 3 and window size 5 to quinoa grains spectra previously corrected by multiplicative scatter correction (MSC; top) or extended multiplicative signal correction (EMSC; bottom).

study may have been a manifestation of our effort to avoid overfitting by consistently selecting the number of latent variables that minimise RMSECV. Moreover, by definition, the coefficient of determination tends to decrease when the range of the dependent variable is lower. The ranges of protein (8.33–11.4% db), fat (5.35–7.78%), carbohydrates (78.5–82.9%) and ashes (2.51–4.11%) essayed from our quinoa samples were narrow in comparison to those from the quinoa samples surveyed in Ferreira et al. (2015) (protein: 11.4–36%, fat: 6.19–15.52%, carbohydrates: 43.6–76.4% and ashes: 3.07–9.15%).

3.4. Influence of SG derivative filters on robustness of chemometric models

Table 3 compiles the SG combinations (m, p, w) leading to the highest predictability within each of the four treatments (i.e., MSC/PLSR, EMSC/PLSR, MSC/CPPLS and EMSC/CPPLS). Although for protein, the same SG filter type ($m = 1$, $p = 2$, $w = 9$) produced the best model's accuracy in the four treatments, this did not necessarily hold for the other dietary constituents (Table 3).

Regardless of the signal correction method and the multivariate algorithm used, SG filtering of quinoa's spectra improved the accuracy of the chemometric models, yet to different degrees: the reduction in RMSECV and RMSEC in the models for moisture (reduction by 1.3–2.6% and 8–14%, respectively), fat (1.5–5.3% and 0.4–1.1%) and ashes (2.1–2.2% and 2.1–10.6%) were all slight in comparison to the considerable reduction in those statistics in the models for protein (8.0–11.9% and 20.5–28.5%) and carbohydrates (8.9–12.4% and 24.2–35.0%). Similarly, SG-filtering improved the correlation statistics of calibration: as before, the increase in R_{CV} and R_C values was slight in the models for moisture (increase by 2.6–5.2% and 0–6.4%, respectively), fat (1.4–5.0% and 0–0.5%) and

Table 2
Accuracy of prediction of NIT chemometric models for quinoa constituents defined by signal correction type (MSC: multiplicative scatter correction, or EMSC: extended multiplicative signal correction) and multivariate algorithm (PLSR: partial least squares regression, or CPPLS: canonical powered partial least squares), as measured by the root mean square errors of cross-validation (RMSECV), calibration (RMSEC) and prediction (RMSEP), and the coefficients of correlation between reference values and those estimated by cross-validation (R_{CV}), calibration (R_C) and prediction (R_P), all of them computed at the minimum number of components. The most robust model for moisture is shown in bold. Please refer to the text.

Proximate composition	Algorithm	Signal correction	Number compo-nents	RMSECV (%)	RMSEC (%)	RMSEP (%)	R_{CV}	R_C	R_P
Moisture	**PLSR**	**MSC**	**8**	**0.575**	**0.480**	**0.592**	**0.576**	**0.732**	**0.596**
		EMSC	5	0.566	0.497	0.615	0.595	0.708	0.551
	CPPLS	MSC	4	0.579	0.607	0.679	0.569	0.507	0.390
		EMSC	4	0.578	0.519	0.601	0.572	0.675	0.579
Protein	PLSR	MSC	8	0.614	0.499	0.549	0.564	0.741	0.738
		EMSC	6	0.584	0.492	0.563	0.619	0.749	0.722
	CPPLS	MSC	4	0.613	0.628	0.629	0.565	0.534	0.634
		EMSC	3	0.588	0.638	0.613	0.611	0.514	0.657
Fat	PLSR	MSC	8	0.326	0.262	0.372	0.718	0.829	0.454
		EMSC	7	0.337	0.266	0.338	0.696	0.822	0.584
	CPPLS	MSC	5	0.325	0.303	0.323	0.719	0.764	0.631
		EMSC	4	0.337	0.308	0.270	0.696	0.753	0.762
Ashes	PLSR	MSC	7	0.231	0.191	0.150	0.742	0.832	0.908
		EMSC	5	0.220	0.190	0.144	0.769	0.833	0.916
	CPPLS	MSC	4	0.233	0.225	0.276	0.737	0.756	0.640
		EMSC	3	0.224	0.217	0.251	0.761	0.776	0.716
Carbohydrates	PLSR	MSC	7	0.612	0.517	0.618	0.689	0.791	0.824
		EMSC	7	0.620	0.501	0.607	0.679	0.806	0.830
	CPPLS	MSC	4	0.595	0.639	0.857	0.710	0.654	0.618
		EMSC	3	0.638	0.662	0.857	0.655	0.621	0.618

ashes (0–1.8% and 1.1–7.1%), whereas the improvement was substantial in the models for protein (13.9–17.3% and 15.6–42.2%) and carbohydrates (8.0–14.5% and 10.8–33%) (percentual differences not shown but calculated from Tables 2 and 3).

The improved RMSECV, RMSEC, R_{CV} and R_C statistics from the models with SG filters for protein and carbohydrates, may be associated to the fact that, for protein and carbohydrates, filtering the spectra led to a higher number of optimal components in the MSC/PLSR (from 8 to 12, and 7 to 12, respectively), EMSC/PLSR (6–10, and 7 to 10), MSC/CPPLS (4–8, and 4 to 10) and EMSC/CPPLS (3–6, and 3 to 8) models. Due to the higher number of components extracted from the SG spectra, the fitting capacity of the protein and carbohydrates models was improved; although the CPPLS algorithm performed better than the PLSR algorithm in the prediction of the test data – as suggested by the differences in RMSEP and R_P. Filtering the spectra with SG largely enhanced the predictive capacity of the models for fat (RMSEP decreased by 1.0–20.4%, and R_P increased by 1.8–24.7%) and ashes (RMSEP decreased by 0.0–30.8%, and R_P increased by 0.0–32.3%), while, as mentioned before, filtering enhanced the prediction performance of the models for protein (RMSEP decreased by 15.8%, and R_P increased by 19.8%), and carbohydrates (RMSEP decreased by 24.8%, and R_P

Table 3
Effect of the best Savitzky-Golay smoothing filter (m: derivative order, p: polynomial order and w:window size) on the accuracy of prediction of NIT chemometric models for quinoa constituents defined by signal correction type (MSC: multiplicative scatter correction, or EMSC: extended multiplicative signal correction) and multivariate algorithm (PLSR: partial least squares regression, or CPPLS: canonical powered partial least squares), as measured by the root mean square errors of cross-validation (RMSECV), calibration (RMSEC) and prediction (RMSEP), and the coefficients of correlation between reference values and those estimated by cross-validation (R_{CV}), calibration (R_C) and prediction (R_P), all of them computed at the minimum number of components. The most robust models for protein, fat, ashes and carbohydrates contents are shown in bold. Please refer to the text.

Proximate composition	Algorithm	Signal correction	Savitzky-Golay			Number comp	RMSECV (%)	RMSEC (%)	RMSEP (%)	R_{CV}	R_C	R_P
			m	p	w							
Moisture	PLSR	MSC	1	3	5	6	0.560	0.441	0.622	0.606	0.779	0.441
		EMSC	1	2	9	4	0.552	0.501	0.629	0.611	0.703	0.501
	CPPLS	MSC	1	2	3	4	0.608	0.521	0.608	0.504	0.673	0.521
		EMSC	1	2	9	3	0.570	0.539	0.586	0.587	0.643	0.539
Protein	PLSR	MSC	1	2	9	12	0.564	0.356	0.592	0.651	0.878	0.685
		EMSC	1	2	9	10	0.527	0.372	0.628	0.705	0.867	0.635
	CPPLS	**MSC**	**1**	**2**	**9**	**8**	**0.564**	**0.486**	**0.529**	**0.651**	**0.757**	**0.760**
		EMSC	1	2	9	6	0.518	0.507	0.635	0.717	0.731	0.625
Fat	PLSR	MSC	2	2	9	5	0.320	0.259	0.344	0.729	0.833	0.565
		EMSC	2	2	9	5	0.319	0.265	0.335	0.732	0.825	0.595
	CPPLS	MSC	1	2	7	4	0.320	0.307	0.257	0.729	0.756	0.787
		EMSC	**2**	**2**	**9**	**3**	**0.327**	**0.310**	**0.248**	**0.716**	**0.751**	**0.804**
Ashes	PLSR	MSC	2	3	9	5	0.226	0.183	0.150	0.756	0.847	0.908
		EMSC	**1**	**3**	**9**	**5**	**0.224**	**0.186**	**0.137**	**0.761**	**0.842**	**0.925**
	CPPLS	MSC	2	3	9	3	0.228	0.202	0.191	0.751	0.810	0.847
		EMSC	1	3	9	4	0.230	0.194	0.175	0.744	0.827	0.873
Carbohydrates	PLSR	MSC	1	2	9	12	0.543	0.358	0.680	0.766	0.906	0.782
		EMSC	1	2	9	10	0.546	0.380	0.747	0.763	0.893	0.728
	CPPLS	**MSC**	**1**	**2**	**9**	**10**	**0.542**	**0.416**	**0.644**	**0.767**	**0.870**	**0.807**
		EMSC	1	2	7	8	0.559	0.444	0.693	0.750	0.776	0.794

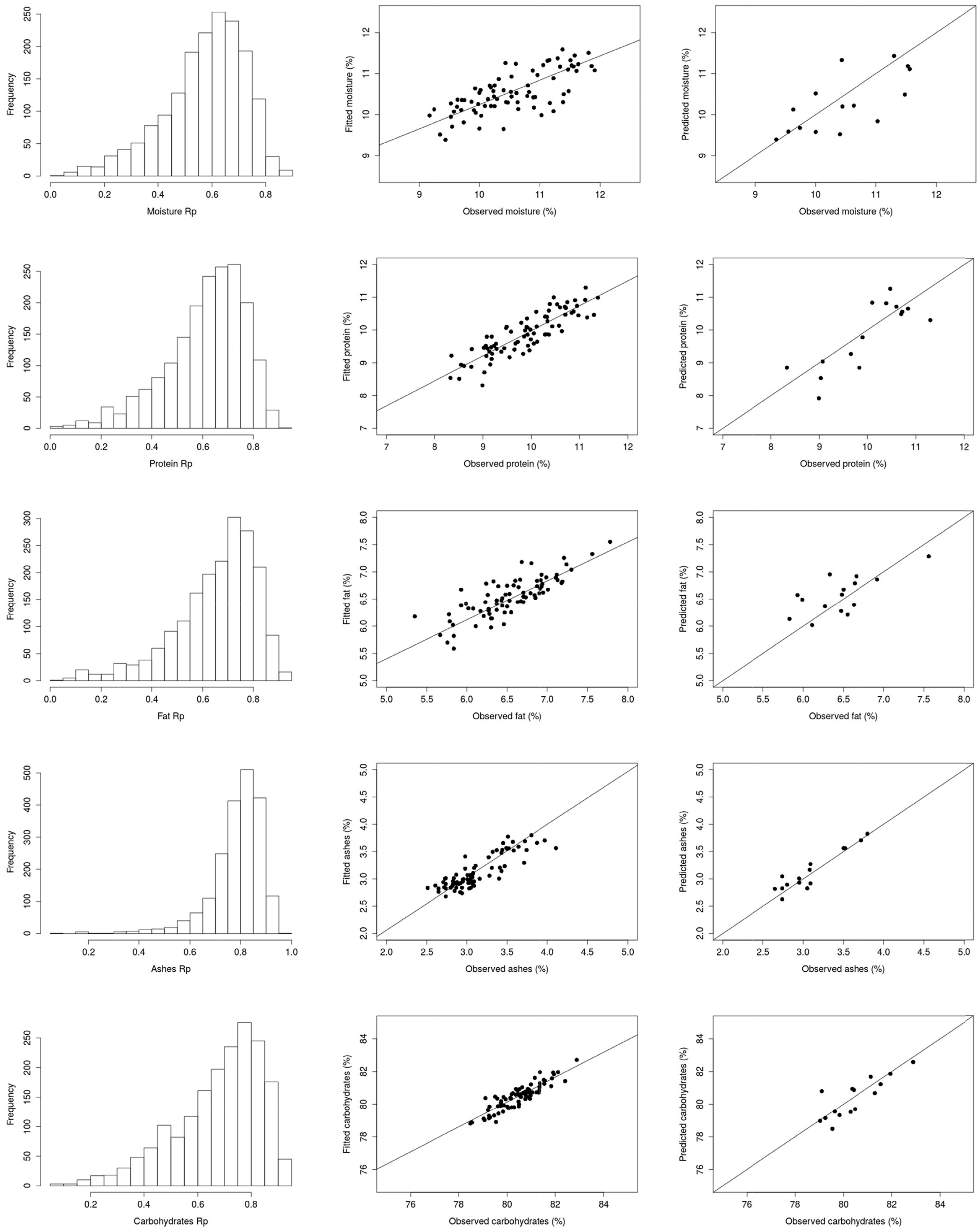

Fig. 3. Prediction performance of NIT chemometric models for moisture, protein, fat, ashes and carbohydrates contents in quinoa grains, as evaluated by the uncertainty about the correlation coefficient of prediction (R_P) built by bootstrapping (left), and the scatter plots between chemical reference values and those fitted to the calibration data set (middle) and predicted using the validation data set (right).

increased by 30.6%) only when CPPLS was used. In the particular case of moisture, only the treatment MSC/CPPLS produced better predictions when spectra were SG-filtered (RMSEP decreased by 10.4%, and R_P increased by 14.1%).

3.5. Validated chemometric models for quinoa's dietary constituents

Taking the four treatments together (Table 3), the models estimating ashes and carbohydrates presented generally the highest predictive capacity, as deduced from the ranges of R_{CV} (0.744–0.761; and 0.750–0.767, respectively) and R_P (0.847–0.925; and 0.728–0.807, respectively). However, the models for protein (R_{CV}: 0.651–0.717; R_P: 0.625–0.760) and fat (R_{CV}: 0.716–0.732; R_P: 0.565–0.804) were of slightly lower predictive performance, while the models for moisture (R_{CV}: 0.504–0.611; R_P: 0.441–0.539) were of fair predictability.

Considering that a good model should bear low values of RMSECV and RMSEP, and high values of R_{CV} and R_P, the final model for each quinoa's constituent was selected among those presented in Tables 2 and 3 For the moisture response, little-to-no gain in prediction performance was attained by SG-filtering the spectra with the many combinations tested. Thus, for this variable, the best model was achieved using a non-filtered spectra treated by MSC and extracting 8 PLSR components, which rendered a prediction CV (RMSEP/mean) of 5.60% and an R_P of 0.596 (other statistics for this model pointed out in bold in Table 2). For the other dietary constituents, better performance was achieved using SG-filtered spectra of window size 9 and first derivative, except for the fat variable which used second derivative. For the NIT determination of ashes, the PLSR algorithm also produced the best model when fitted to EMSC-treated spectra. The 5 optimal latent variables extracted yielded on the test data a CV of 4.38% and R_P of 0.925. For the protein, fat and carbohydrates variables, the CPPLS multivariate algorithm performed better: whilst the best predictability of protein (CV = 5.35% and R_P = 0.760) was achieved by extracting 8 components from MSC-treated spectra, the best model for carbohydrates was produced by extracting 10 components from MSC-treated spectra (CV = 0.80% and R_P = 0.807). With a CV = 3.79% and R_P = 0.804, fat could be estimated by a CPPLS model produced from a EMSC-treated spectra with only 3 latent variables.

Finally, in order to further characterise the prediction performance of each of the final models, uncertainty about the correlation coefficient of prediction (R_P) was built by bootstrapping. At each of the 1000 iterations, a new 80% calibration/20% validation data partition was randomly obtained, the chosen model was fitted to the calibration data with the pre-determined number of components, and R_P was extracted from the test data. The histograms of R_P built for each of the final models (Fig. 3, left) show that the NIT model for estimating ashes had the lowest uncertainty (i.e., narrow spread) about R_P, and therefore was the most robust chemometric model. The wider spread of the R_P histogram for moisture corroborated that, among the five dietary constituents studied, the model for moisture presented the lowest precision. The degree of fitting and predictability of the final models can be appreciated from the scatter plots between the reference values and those fitted (Fig. 3, middle) and predicted (Fig. 3, right) from the NIT calibration models. The best agreement between observed and predicted values was observed for ashes and carbohydrates; although, as a whole, the degree of dispersion in the predictions is acceptable, bearing in mind that chemical analyses also have associated errors.

4. Conclusions

Regardless of the multivariate algorithm used, light scattering correction of quinoa grains' NIT spectra by EMSC consistently led to proximate composition models of better cross-validation statistics – except for fat and carbohydrates – than those produced by MSC-treated spectra. Both EMSC, as opposed to MSC; and CPPLS, as opposed to PLSR, led to fewer optimal components. When spectra were treated by different types of SG filters, the optimal latent variables reduced correspondingly in each of the four treatments (i.e., MSC/PLSR, EMSC/PLSR, MSC/CPPLS, EMSC/CPPLS), except for the models predicting protein and carbohydrates, in which the behaviour was the opposite. In addition, smoothing the quinoa's spectra enhanced the accuracy of the models for fat, ashes, and particularly for protein and carbohydrates, while improving also the prediction performance, particularly, for fat and ashes determination. Although the most robust models could be developed for ashes (SG-filtered EMSC/PLSR: 90% confidence interval for RMSEP [0.376–0.512] as determined by bootstrap) and carbohydrates (SG-filtered MSC/CPPLS: 90% CI RMSEP: [0.651–0.901]), the predictability was still acceptable for the other dietary constituents; namely, protein (SG-filtered MSC/CPPLS: 90% CI RMSEP: [0.650–0.852]), fat (SG-filtered EMSC/CPPLS: 90% CI RMSEP: [0.478–0.654]) and moisture (non-filtered EMSC/PLSR: 90% CI RMSEP: [0.658–0.833]). Thus, in this study, satisfactory predictions of the dietary constituents of quinoa grains could be achieved by using NIT technology. The main advantages of the technique are the rapid determination for routine analysis, the reduced costs and absence of sample preparation and waste generation.

Acknowledgments

Mr. Encina-Zelada acknowledges the financial aid provided by the Peruvian National Programme of Scholarships and Student Loans (PRONABEC) in the mode of PhD grants (Presidente de La República Grant Number 183308). Dr. Gonzales-Barron wishes to acknowledge the financial support provided by the Portuguese Foundation for Science and Technology (FCT) through the award of a five-year Investigator Fellowship (IF) in the mode of Development Grants (IF/00570).

References

AOAC. (2000). In W. Horwitz (Ed.), *Official methods of analysis of the association of analytical Chemists international* (17th ed.). Gaithersburg, MD, USA: AOAC International.

Büchman, N. B., Josefsson, H., & Cowe, I. A. (2001). Performance of European artificial neural network (ANN) calibrations for moisture and protein in cereals using the Danish near infrared transmission (NIT) network. *Cereal Chemistry, 78*(5), 572–577.

Cantor, S. L., Hoag, S. W., Ellison, C. D., Khan, M. A., & Lyon, R. C. (2011). NIR spectroscopy applications in the development of a compacted multiparticulate system for modified release. *Journal of the American Association of Pharmaceutical Scientists, 12*(1), 262–278.

Ferreira, D. S., Pallone, J. A. L., & Poppi, R. J. (2015). Direct analysis of the main chemical constituents in *Chenopodium quinoa* grain using Fourier transform near-infrared spectroscopy. *Food Control, 48*, 91–95.

González-Martín, M. I., Moncada, G. W., Fischer, S., & Escuredo, O. (2014). Chemical characteristics and mineral composition of quinoa by near-infrared spectroscopy. *Journal of the Science of Food and Agriculture, 94*(5), 876–881.

Indahl, U. G., Liland, K. H., & Næs, T. (2009). Canonical partial least squares -a unified PLS approach to classification and regression problems. *Journal of Chemometrics, 23*, 495–504.

Jancurová, M., Minarovicová, L., & Dandar, A. (2009). Quinoa - a review. *Czech Journal of Food Sciences, 27*(2), 71–79.

Liland, K. H. (2016). *Extended multiplicative signal correction. Package "EMSC"*. Date 2016-04-24. Repository CRAN. Available online at: https://cran.r-project.org/web/packages/EMSC/index.html (Accessed: 16.05.2016).

Maleki, M. R., Mouazen, A. M., Ramon, H., & De Baerdemaeker, J. (2007). Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems Engineering, 96*(3), 427–433.

Martens, H., & Stark, E. (1991). Extended multiplicative signal orrection and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis, 9*(8), 625–635.

Mevik, B. H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software, 18*(2), 1–24.

Mevik, B. H., Wehrens, R., & Liland, K. H. (2015). *Pls: Partial least squares and principal component regression*. R package version 2.5-0. Available online at: https://cran.r-project.org/web/packages/pls/ (Accessed: 16.05.2016).

Miralbés, C. (2004). Quality control in the milling industry using near infrared transmittance spectroscopy. *Food Chemistry, 88*(4), 621–628.

Moghimi, A., Aghkhani, M. H., Sazgarnia, A., & Sarmad, M. (2010). Vis/NIR spectroscopy and chemometrics for the prediction of soluble solids content and acidity (pH) of kiwifruit. *Biosystems Engineering, 106*(3), 295–302.

Panero, P. S., Panero, F. S., Panero, J. S., & Silva, H. E. B. (2013). Application of extended multiplicative signal correction to short-wavelength near infrared spectra of moisture in marzipan. *Journal of Data Analysis and Information Processing, 1*(3), 30–34.

Pojić, M., Mastilović, J., Pestorić, M., & Radusin, T. (2008). The ensuring of measurements for cereal quality determination. *Food Processing, Quality and Safety, 35*(1), 11–18.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Stastistical Computing. Available online at: http://www.R-project.org/ (Accessed: 04.02.2016).

Repo-Carrasco-Valencia, R., Hellström, J. K., Pihlava, J. M., & Mattila, P. H. (2010). Flavonoids and other phenolic compounds in Andean indigenous grains: Quinoa (*Chenopodium quinoa*), kañiwa (*Chenopodium pallidicaule*) and kiwicha (*Amaranthus caudatus*). *Food Chemistry, 120*(1), 128–133.

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639.

Stevens, A., & Ramirez-Lopez, L. (2013). *An introduction to the prospectr package*. Vignette R package version 0.1.3. Available online at: https://github.com/antoinestevens/prospectr (Accessed: 16.05.2016).

Vega-Gálvez, A., Miranda, M., Vergara, J., Uribe, E., Puente, L., & Martínez, E. (2010). Nutrition facts and functional potential of quinoa (*Chenopodium quinoa* willd.), an ancient andean grain: A review. *Journal of the Science of Food and Agriculture, 90*(15), 2541–2547.

Wold, H., Martens, H., & Wold, S. (1983). The multivariate calibration method in chemistry solved by the PLS method. In A. Ruhe, & B. Kågström (Eds.), *Proceedings of the conference of matrix pencils, lecture notes in mathematics* (pp. 286–293). Heidelberg: Springer Verlag.