LWT - Food Science and Technology 151 (2021) 112192

Contents lists available at ScienceDirect

LWT

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

Online NIRS analysis for the routine assessment of the nitrate content in spinach plants in the processing industry using linear and non-linear methods

Miguel Vega-Castellote [a], Dolores Pérez-Marín [b,**], Irina Torres [a], María-Teresa Sánchez [a,*]

[a] Department of Bromatology and Food Technology, University of Cordoba, Rabanales Campus, 14071, Córdoba, Spain
[b] Department of Animal Production, University of Cordoba, Rabanales Campus, 14071, Córdoba, Spain

ARTICLE INFO

Keywords:
Quantitative models
MPLS algorithm
LOCAL algorithm
Sample variability

ABSTRACT

This study aimed to assess the robustness of the NIRS models developed following different strategies for the routine prediction of nitrate content in spinach plants using an online FT-NIR spectrophotometer. To achieve this, 516 spinach plants from different cultivars, harvest dates, orchards and seasons, were used. Two strategies were followed to make up the calibration and validation sets; the first included in the calibration set those samples belonging to the 2018 and 2019 harvesting seasons, while the second also included in this set part of the population of the 2020 harvesting season. Modified partial least squares quantitative models were initially developed and externally validated. In view of the results and to obtain significant improvements, a non-linear regression technique (the LOCAL algorithm) was applied. The models developed using the non-linear regression technique and considering the greatest possible variability in the training set (samples belonging to 2018, 2019 and 2020 harvesting seasons) reported the best prediction results ($R^2_p = 0.60$; SEP = 758 mg/kg), which enabled to classify the product in the main categories or classes established by the official regulations, according to its commercial destination.

1. Introduction

Decision-making regarding the postharvest management of horticultural products should be based on information related to, among other factors, the safety of the different measurement techniques (Walsh, McGlone, & Han, 2020). In spinach, it is important to check the nitrate content, which is limited by European Union (EU) food safety regulations, since high concentrations can have detrimental effects on human health (Jaworska, 2005), and it is therefore necessary to develop fast, non-destructive, and high throughput analysis systems which can be implemented at an industrial level.

Near infrared spectroscopy (NIRS) in combination with multivariate analysis methods enables to meet the industry requirements, since it offers the possibility to assess non-destructively the nitrate content in spinach plants on the industrial sorting lines and at a reduced cost (Entrenas, Pérez-Marín, Torres, Garrido-Varo, & Sánchez, 2020; Torres, Sánchez, Entrenas, Garrido-Varo, & Pérez-Marín, 2020). The NIRS calibration studies developed for the prediction of the nitrate content in spinach plants have often used linear regression techniques such as Partial Least Squares (PLS) –which has demonstrated its potential ability to estimate the nitrate content in these plants– to develop the so-called global equations (Entrenas et al., 2020; Mahanti, Chakraborty, Kotwaliwale, & Vishwakarma, 2020; Pérez-Marín, Torres, Entrenas, Vega, & Sánchez, 2019; Torres, Sánchez, & Pérez-Marín, 2020; Torres, Sánchez, Vega-Castellote, Luqui-Muñoz, & Pérez-Marín, 2021), in which all the samples belonging to the calibration set are used to build the prediction model (Pérez-Marín, Garrido-Varo, & Guerrero, 2005), and have provided model performance statistics using data belonging to a single harvesting season.

According to Peirs, Tirry, Verlinden, Darius, and Nicolaï (2003) and Subedi, Walsh, and Hopkins (2012), the accuracy of the NIRS model relies to a large extent upon the spectral variability included in the calibration set, and the cultivar, growing season and growing region are the main factors contributing to this variability. Thus, Walsh et al. (2020) recommended considering a minimum of three harvesting seasons to develop robust NIRS prediction models for agricultural products.

* Corresponding author.
** Corresponding author
E-mail addresses: dcperez@uco.es (D. Pérez-Marín), teresa.sanchez@uco.es (M.-T. Sánchez).

https://doi.org/10.1016/j.lwt.2021.112192
Received 10 May 2021; Received in revised form 9 July 2021; Accepted 21 July 2021
Available online 23 July 2021

However, according to Berzagui, Shenk, and Westerhaus (2000), although the use of large number of samples in the calibration databases would properly represent the population variability and increase the model robustness, it often reduces the accuracy of the prediction. In addition, calibrations with large databases using PLS regression would need a higher number of terms to explain their variability and, thereby, more complex models would be obtained (Shenk, Westerhaus, & Berzaghi, 1997). Furthermore, Dambergs, Cozzolino, Cynkar, Janik, and Gishen (2006) and Pérez-Marín, Fearn, Guerrero, and Garrido-Varo (2012) pointed out that wide ranges of variation of the parameter under consideration, in addition to high heterogeneity among the samples, could be indicators of a need for more sophisticated algorithms to develop the prediction models.

This issue could be addressed using local regressions, which are based on selecting samples from a large dataset to develop specific calibrations for each of those samples to be predicted. One of these methods is the LOCAL algorithm, proposed by Shenk et al. (1997), which both enables us to cover the variability of the original database and offers the accuracy obtainable with specific calibrations (Pérez-Marín, Garrido-Varo, & Guerrero, 2007; Saeys, Do Trong, Van Beers, & Nicolaï, 2019). This algorithm allows us to manage non-linear data and has been used successfully in several applications in recent years involving vegetal products such as grapes (Dambergs et al., 2006), nectarines (Sánchez, De la Haba, Guerrero, Garrido-Varo & Pérez-Marín, 2011), citrus fruit (Torres, Pérez-Marín, De la Haba & Sánchez, 2017; Torres, Sánchez, De la Haba, & Pérez-Marín, 2019) or almonds (Vega-Castellote, Pérez-Marín, Torres, Moreno-Rojas, & Sánchez, 2021).

The aim of this study was, firstly, to evaluate the robustness of the models developed for the prediction of nitrate content in spinach plants analysed with the Fourier transformed near infrared (FT-NIR) Matrix-F instrument, simulating industrial selection and classification processes, and using the MPLS (Modified Partial Least Squares) linear regression with a calibration set composed of plants belonging to three different seasons (2018, 2019 and 2020) and growing areas (different orchards in the provinces of Cordoba and Seville, Spain). Secondly, depending on the results obtained, i.e. if the predictive capacity of the nitrate prediction models did not increase despite increasing the variability of the calibration set, we considered evaluating the LOCAL algorithm which allowed to model the non-linearity of the data.

2. Material and methods

2.1. Sampling and reference data

In this study, a total of 516 spinach plants (*Spinacia oleracea* L. cv. '*Armónica*', '*PV-1194*', '*Baboon*', '*Bandicoot*', '*Gorilla*' and '*Solomon*') grown outdoors on different orchards in the provinces of Cordoba and Seville (Spain) throughout the 2018 (N = 195 plants), 2019 (N = 228 plants), and 2020 (N = 93 plants) seasons were used. These plants were harvested manually during the months of January and March of these years.

Nitrate content (mg/kg) was measured according to Pérez-Marín et al. (2019). First, between 4 and 10 spinach leaves were cut into small pieces, liquefied and then filtered. Next, the nitrate concentration in the extraction liquid was measured using a RQflex reflectometer (Merck, Darmstadt Germany). The analytical measurements were carried out in duplicate, and the standard error of laboratory (SEL) was calculated from those replicates.

2.2. NIR spectrum acquisition

The spectral data from the spinach plants was collected using an online FT-NIR spectrophotometer, the Matrix-F (Bruker Optik GmbH, Ettlingen, Germany). This instrument consists of a detection head with two NIR light sources, which illuminate a sampling area of around 154 cm^2. This detection head was attached to the spectrophotometer via a 5 m length fibre optic probe. The spectra were collected in reflectance mode in the spectral range from 4000 to 12,000 cm^{-1} (834–2502.40 nm), with a resolution of 16 cm^{-1} (1.61 nm). The instrument's performance was checked every 30 min by collecting an internal white reference.

In this study, the spectral information of the spinach plants was obtained by simulating the online analysis carried out in the industry. To achieve this, the system featured a conveyor belt to move the sample. The samples were therefore analysed in dynamic mode –i.e. with the conveyor belt in motion (4.77 cm/s)– with 16 scans taken per spectrum. The scan time per sample was of 8 s, and 2 spectra were taken per plant. A mean spectrum per sample was obtained by averaging the 2 spectra taken per plant.

2.3. Data pre-processing and definition of the calibration and validation sets for the different strategies

In this study, the spectral data were subjected to pre-processing and chemometric treatments using WinISI II software package version 1.50 (Intrasoft International LLC, Port Matilda, PA, USA) and Matlab software version 2019a (The Mathworks, Inc., Natick, MA, USA).

Firstly, the areas of the spectrum where the signal to noise ratio was degraded were eliminated. To select the optimum spectral range for the Matrix-F instrument, the 1,1,1,1 derivation treatment was applied (being the first digit the number of the derivative, the second the gap over which the derivative is calculated, the third the number of data points in a running average or smoothing, and the fourth the second smoothing) without scatter correction, which allows to highlight the areas of the spectrum where the signal/noise ratio is degraded (Hruschka, 2001).

Next, to define the calibration and validation sample sets, two strategies were followed.

- Strategy I. The set of samples belonging to the 2018 and 2019 harvesting seasons were used to build the calibration set (C1, 423 samples). The best model was later externally validated using the 2020 season samples (V1, 93 samples).
- Strategy II. Part of the variability of the 2020 harvesting season was included in the calibration set, as is recommended in routine analysis when NIRS is used, by adding the first 30 samples of the 2020 season to the C1 set (C2, 453 samples). In this case, the external validation set was built using the remaining 63 samples from the 2020 season (V2).

A Principal Components Analysis (PCA) was performed on the C1 and C2 calibration sets applying the CENTER algorithm, which calculates the centre of the population and the Mahalanobis distance (GH) of each sample to that centre. Those samples with a statistical value of GH greater than 4 could be considered as spectral outliers and were studied in detail to make the decision of removing them from their respective sample set (Shenk & Westerhaus, 1995a). This algorithm was applied using a combination of the Standard Normal Variate (SNV) and Detrending (DT) mathematical pre-treatments to remove scatter interferences, together with the 1,5,5,1 gap-segment first derivative treatment (Barnes, Dhanoa, & Lister, 1989; Shenk & Westerhaus, 1995b). After studying the potential outliers, the C1 and C2 calibration groups were renamed as E1 and E2, respectively.

2.4. Prediction of the nitrate content using a linear regression algorithm

To predict the nitrate content, MPLS regression algorithm was applied to the E1 and E2 calibration sets (Shenk & Westerhaus, 1995a). The regression equations were obtained using the combination of SNV and DT (Barnes et al., 1989) for scatter correction and testing different order and gap-segment derivative treatments: i.e., 1,5,5,1 and 2,5,5,1 (Shenk & Westerhaus, 1995b). Cross-validation was performed using 4

groups, and the best equations were selected considering the coefficient of determination of cross-validation (R^2_{cv}), standard error of cross-validation (SECV) and the residual predictive deviation for cross-validation (RPD_{cv}), which is a statistic calculated as the ratio of the standard deviation (SD) of the reference data to the SECV and which enables us to standardize the SECV and thereby to compare results obtained with sets of different means (Shenk & Westerhaus, 1996; Williams, 2001). These calibration models developed using the E1 and E2 sets were externally validated using the V1 and V2 sets, respectively. The validation procedure was carried out following the protocol outlined by Windham Dhanoa, & Lister, based on the following statistics: the coefficient of determination for external validation (R^2_p), the standard error of prediction (SEP), the standard error of prediction corrected for bias ($SEP_{(c)}$) and the bias.

Initially, the structure and variability of the validation sets were not studied, since a routine industrial case study was simulated in this work and, consequently, the prediction results obtained with the MPLS algorithm were evaluated using control reliability statistics based on the spectral distances of the sample to the centre of the calibration population (GH), and the spectral distance between the sample and neighbouring or similar samples (the neighbourhood Mahalanobis distance (NH)) (Pérez-Marín, Garrido-Varo, Riccioli, & Fearn, 2018). The limits for the control statistics for the routine online analysis was set according to Torres et al. (2021), who established a maximum value of GH = 4 and NH = 1 for spinach plants. These GH and NH values were studied for all the samples belonging to the V1 and V2 sets and, more specifically, for those predicted samples showing high residual predictive values i.e., samples showing a Student's t value above 2.5 (Jerome & Workman, 2008), with the Student's t value calculated as the ratio of the residual value for a given sample to the SEP.

The results obtained with this linear algorithm led to the development of an in-depth study of the SECV obtained when the models were devised using the E1 and E2 sets, in order to determine whether this error was constant throughout the nitrate content range. The coefficient of variation (CV) was also studied by comparing the SECV to the mean value of the reference data (Williams, 2001). After that, new MPLS models were created by removing from the E1 and E2 sets those samples showing the highest CV, in order to evaluate the results of the developing models without those values of the nitrates range showing the greatest relative SECV compared to their mean reference data value. In this study, the limit for the CV was set at 100%, in which the SECV value equals the mean value of the reference data.

2.5. Study of the performance of models developed using a non-linear regression technique

The LOCAL algorithm was performed using the E1 and V1 and E2 and V2 sets in order to evaluate the prediction capabilities of this non-linear regression technique. Using the LOCAL algorithm, samples in the validation sets (V1 and V2) were predicted by selecting those samples in the calibration sets (E1 and E2) with a similar spectrum to the one being analysed. This selection is controlled by means of the correlation coefficient value between the spectrum of the sample to be predicted and the spectra in the product database (Shenk et al., 1997). The number of calibration samples (k) and the number of PLS factors (l) were assessed in order to optimize the LOCAL algorithm. The 'k' value was set from 60 to 200 in steps of 20, and 'l' from 6 to 16 in steps of 2. This gave a factorial design of 8 × 6 or 48 runs. Shenk et al. (1997) reported that the accuracy of LOCAL predictions could be improved by excluding the prediction values generated with the first few PLS factors and, consequently, it was agreed that the first three PLS factors should be removed. The same pre-treatments as for MPLS regression were used. The performance of the LOCAL algorithm was assessed considering the R^2_p, SEP, $SEP_{(c)}$ and bias. The prediction results obtained with the LOCAL algorithm were also evaluated using the GH and NH control reliability statistics, as explained previously for the linear regression algorithm.

Furthermore, the SEP values for the predictive models obtained using the MPLS and LOCAL algorithms were compared using Fisher's F test (Naes, Isaksson, Fearn, & Davies, 2002). Values for F were calculated as:

$$F = \frac{(SEP_2)^2}{(SEP_1)^2}$$

where SEP_1 and SEP_2 are the standard errors of prediction and $SEP_1 < SEP_2$. F is compared to $F_{critical}$ $(1 - P, n_1 - 1, n_2 - 1)$, as read from the table, with $P = 0.05$, and n_1 is the number of times the measurement is repeated with method 1, while n_2 is the number of times the measurement is repeated with method 2. If F is higher than $F_{critical}$, the two SEP values are significantly different.

3. Results and discussion

3.1. Selection of the optimal NIR spectral region

To obtain sample-representative and high-quality spectra, the Matrix-F range of work was evaluated, and the optimal spectral work region was selected. This aspect is essential to obtain robust NIRS models. It was observed that the regions between 834-1475 nm and 2403–2502 nm (Fig. 1) showed high levels of noise. The spectral signal in the Matrix-F instrument is transmitted via fibre optics, which commonly produce a loss of signal quality on extreme wavelengths (Garrido-Varo, Sánchez-Bonilla, Maroto-Molina, Riccioli, & Pérez-Marín, 2018). Furthermore, the initial spectral region showed limited spectral information. These regions were therefore eliminated.

3.2. Characterization of the calibration and validation sample sets and identification of spectral outliers

After applying the CENTER algorithm to the calibration sets, a total of 9 and 10 samples showed GH value > 4 for the C1 and C2 groups, respectively. In both cases, only 3 samples were eliminated –the same three samples in C1 and C2 groups– which showed extreme GH values when the CENTER algorithm was applied to C1 (GH = 9.77, 12.55 and 14.17) and C2 (GH = 10.36, 12.82 and 14.30). One of these samples showed anomalies in the spectrum curve, which could be attributed to an error in the spectrum acquisition process. Next, the C1 and C2 sets were renamed as E1 (N = 420 samples) and E2 (N = 450 samples).

For strategies I and II, the calibration sets covered the variability of the validation sets (Fig. 2), with these sets of samples showing similar values for nitrate content (Table 1). Pérez-Marín et al. (2005) highlighted the importance of a correct selection of those samples included in the calibration set, since this has a major effect on the precision and accuracy of the calibrations performed. The coefficient of variation (CV) for nitrate content, in all cases, showed values over 65%, since this parameter is highly dependent on several factors such as the fertilization, growing stage, soil characteristics, cultivar and climatological conditions throughout the growing period (Proietti, Moscatello, Giacomelli, & Battistelli, 2013; Colla, Kim, Kyriacou, & Rouphael, 2018).

3.3. Prediction of safety parameter using a linear regression technique

The cross-validation results obtained in this study (Table 2) showed that the best calibration models obtained for the prediction of nitrate content in spinach plants for strategies I and II would allow to classify samples as showing high and low values of this parameter (Shenk & Westerhaus, 1996; Williams, 2001). The best calibration models developed using the E1 and E2 sets to predict this parameter were validated using the V1 and V2 sets for the two strategies tested (Fig. 3). Only one sample in V1 and in V2 –the same for both sets– showed a NH value higher than the limit ($NH_{V1} = 1.05$ and $NH_{V2} = 1.06$). However, this sample did not show GH values > 4 nor T values > 2.5 and,

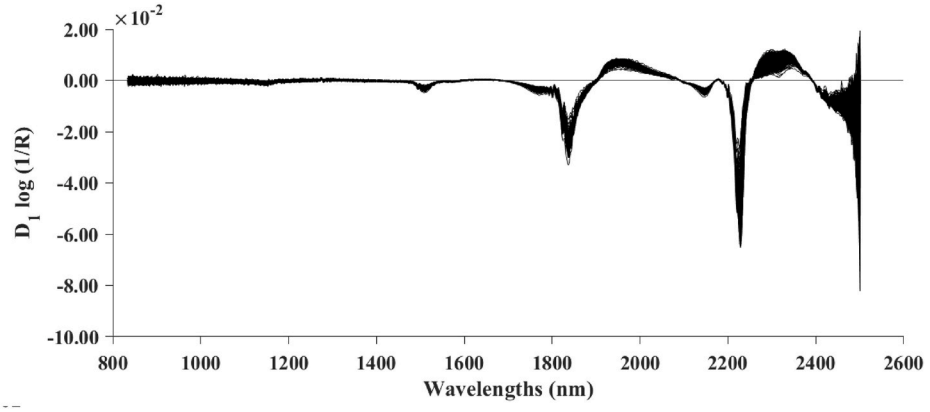

Fig. 1. First derivative spectral values for spinach plants analysed using the Matrix-F online instrument.

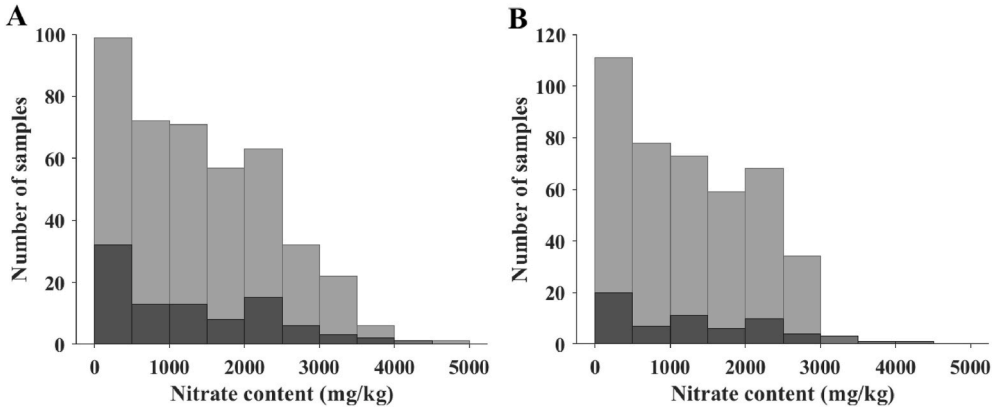

Fig. 2. Distribution of the samples belonging to the E1 and V1 (A – Strategy I) and to the E2 and V2 (B – Strategy II) calibration and validations sets for nitrate content. – Calibration samples; – Validation samples.

Table 1
Characterization of the calibration (E1 and E2) and validation (V1 and V2) sets for nitrate content following strategies I and II.

Parameter	Strategy	Set	Group	N	Range	Mean	SD	CV (%)
Nitrate content (mg/kg)	I	Calibration	E1	420	41–4934	1406	985	70
		Validation	V1	93	80–4249	1299	1058	81
	II	Calibration	E2	450	41–4934	1386	991	66
		Validation	V2	63	80–4249	1386	1057	76

N – Number of samples; SD – Standard deviation; CV – Coefficient of variation.

Table 2
Calibration statistics for the prediction of the nitrate content in spinach plants using the modified partial least squares –MPLS– linear regression technique and the 1,5,5,1 math treatment.

Parameter	Strategy	Group	N	Range	Mean	SD	LV	R^2_{cv}	SECV	RPD_{cv}	SEL
Nitrate content (mg/kg)	I	E1	410	41–3845	1367	937	7	0.43	707	1.32	140
	II	E2	435	41–3817	1346	947	9	0.45	700	1.35	

SD – Standard deviation; LV – Number of latent variables; R^2_{cv} – Coefficient of determination of cross-validation; SECV – Standard error of cross-validation; RPD_{cv} – Residual predictive deviation for cross-validation; SEL – Standard error of laboratory.

consequently, it was not removed from the validation sets.

No significant differences were found in terms of the SEP ($P > 0.05$) between the two strategies tested. One sample belonging to the V2 set showed a NIR predicted negative value, and thus, this value was shown as zero (Fig. 3B). For strategies I and II, the values for R^2_p did not meet the requirements set by Windham, Mertens, and Barton (1989) ($R^2_p > 0.6$). However, the $SEP_{(c)}$ and the bias showed values below the recommended limits by the protocol mentioned above. Consequently, these equations can be taken as an initial approximation to the online measurement of the nitrate content in spinach plants in the industry sorting lines.

According to Shenk and Westerhaus (1996) and Williams (2001), models showing SEP values under 2 × SEL can be considered as models with a high predictive capacity. In our research, the SEP values were above 2 × SEL for both strategies assayed (Fig. 3). However, it should be taken into account that the limit recommended by the scientific literature refers to different NIR analysis conditions which do not involve the use of perishable products like spinach plants, but pre-dried and ground

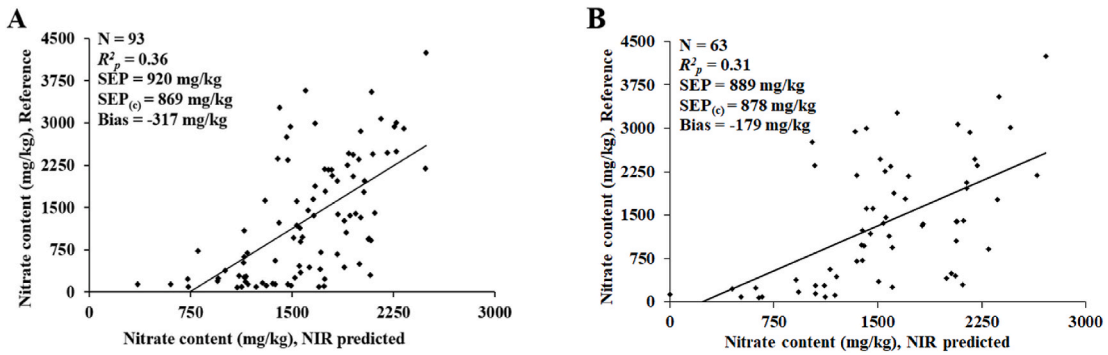

Fig. 3. Reference and near infrared (NIR) predicted values for nitrate content for strategy I (A) and strategy II (B) using the modified partial least squares (MPLS) algorithm. N – Number of samples of the validation set; R^2_p – Coefficient of determination for prediction; SEP – Standard error of prediction; $SEP_{(c)}$ – Standard error of prediction corrected for bias.

samples.

It is important to note that only one previous study was found in the literature which aimed to develop an online NIRS application to predict the nitrate content in spinach plants in the food industry. That study, carried out by Entrenas et al. (2020), involved a total of just N = 195 samples from a single harvesting season ($R^2_p = 0.41$; SEP = 663 mg/kg), which can be considered as a feasibility study but with an insufficient number of samples to obtain robust NIRS calibration models. In fact, the model developed using MPLS regression for the prediction of nitrate content cannot be considered as robust either, although three harvesting seasons were included. Consequently, new strategies to enhance the obtained results were developed and, to achieve this, the SECV obtained for strategies I and II were studied (Fig. 4). The extreme ends of the nitrate content range showed, for both strategies, the highest SECV values. Nevertheless, the greatest relative errors expressed as the CV were found for the lowest values of the nitrate content range.

To develop the new models, those samples with a nitrate content under 600 mg/kg were removed from the calibration sets, both for strategies I and II, since the samples in the range 0–600 mg/kg presented CV values over 100%. The new models developed (Table 3) allowed us to reduce the SECV values, but they did not enhance the prediction capacity of those ones developed using the whole range of nitrates in terms of the RPD_{CV}.

3.4. Optimization of settings, development of predictive models using the LOCAL algorithm and comparison between the best models developed using the linear and non-linear regression techniques

Taking into account the previous results obtained with linear prediction models, where the robustness and prediction ability did not improve despite increasing the calibration set variability using samples collected throughout three harvesting seasons, it was decided to test another strategy to develop the models, using in this case the LOCAL algorithm.

The application of the LOCAL algorithm using the best mathematical pre-treatments and the 48 runs carried out (Table 4) showed that for the nitrate content, the optimum number of calibration samples (k) was 140 and 180 for strategies I and II, respectively, since adding more samples would make the SEP greater (Fig. 5). In addition, the lowest SEP value for the different PLS factors used was obtained for 'l' = 16 for both strategies. Five and six samples belonging to the V1 and V2 sets, respectively, showed a NH value greater than 1. None of these samples presented a GH value over the limit and only one of them –belonging to the V2 set– showed a T value > 2.5, which was consequently removed from that set. Moreover, this sample presented an extreme nitrate content value close to 3000 mg/kg, which can account for the NH value > 1, since its reference value indicates that it is a sample with a low representation in the calibration set (Fig. 2B).

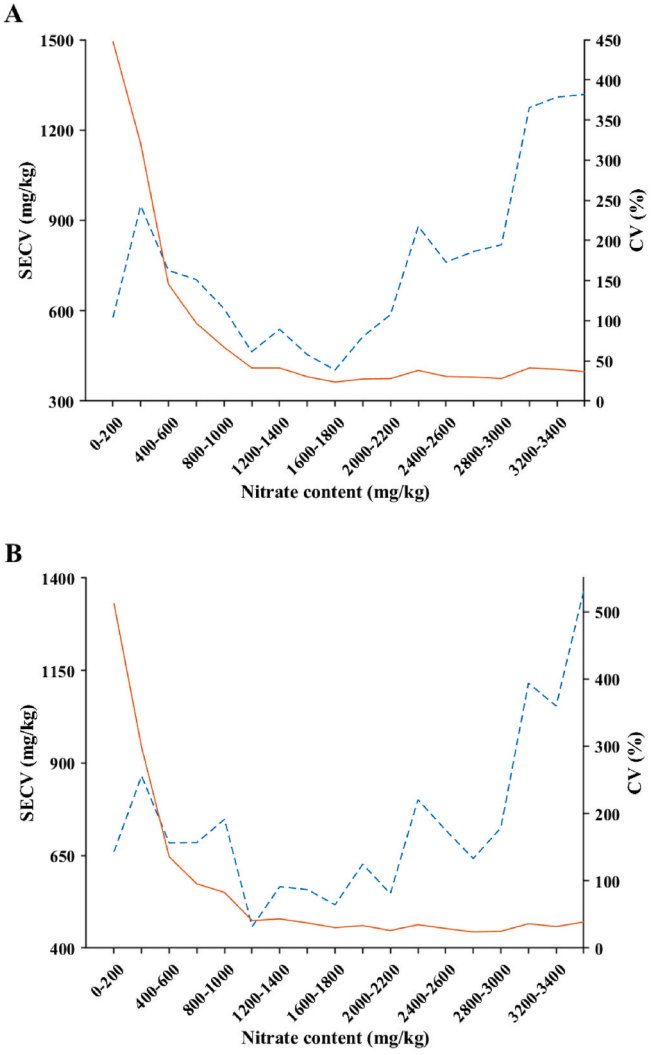

Fig. 4. Standard error of cross-validation (SECV) and coefficient of variation (CV) values obtained throughout the nitrate content range when calibration models were developed using the modified partial least squares (MPLS) algorithm following strategies I (A) and II (B). — – Coefficient of variation (CV); - - - – Standard error of cross-validation (SECV).

According to Shenk and Westerhaus (1996) and Williams (2001), the results obtained for the prediction of nitrate content using the LOCAL algorithm indicate that the model devised following the first strategy

Table 3
Calibration statistics for the prediction of the nitrate content using samples showing nitrate content over 600 mg/kg and the modified partial least squares –MPLS– linear regression technique and the 1,5,5,1 math treatment.

Parameter	Strategy	N	Range	Mean	SD	LV	R^2_{cv}	SECV	RPD_{cv}
Nitrate content (mg/kg)	I	306	630–3845	1779	769	4	0.16	704	1.09
	II	322	623–3793	1772	768	7	0.22	679	1.13

SD – Standard deviation; LV – Number of latent variables; R^2_{cv} – Coefficient of determination of cross-validation; SECV – Standard error of cross-validation; RPD_{cv} – Residual predictive deviation for cross-validation.

Table 4
Validation statistics for the prediction of the nitrate content in spinach plants using the LOCAL algorithm and the 1,5,5,1 math treatment.

Parameter	Strategy	Calibration samples (*k*)	N	Factors (*l*)	R^2_p	SEP	$SEP_{(c)}$	Bias	RPD_p
Nitrate content (mg/kg)	I	140	93	16 (−3)	0.38	940	897	−358	1.12
	II	180	62	16 (−3)	0.60	758	738	−199	1.39

N – Number of samples; R^2_p – Coefficient of determination for prediction; SEP – Standard error of prediction; $SEP_{(c)}$: Standard error of prediction corrected for bias; RPD_p – residual predictive deviation for prediction.

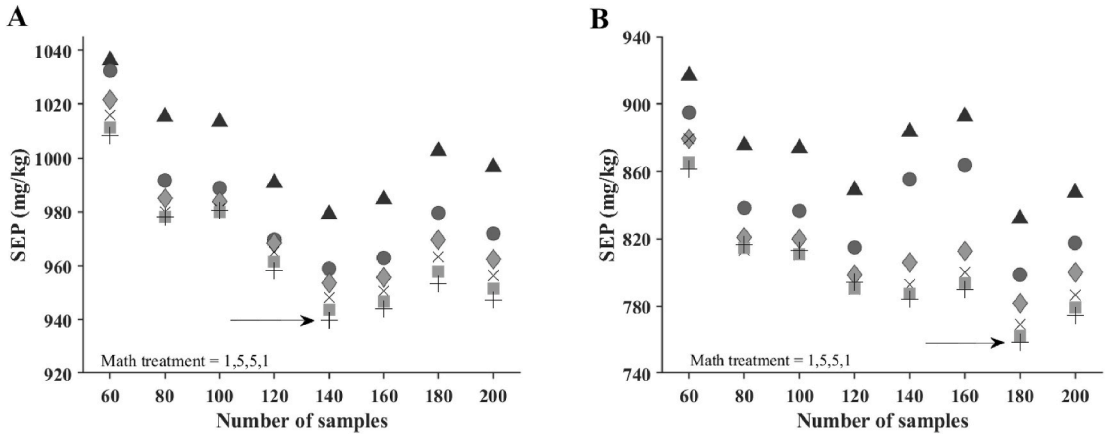

Fig. 5. Standard error of prediction (SEP) values obtained for the prediction of the nitrate content in spinach plants using the LOCAL algorithm using the strategies I (A) and II (B). ▲ – 6 factors; ● – 8 factors; ◆ – 10 factors; × – 12 factors; ■ – 14 factors; + – 16 factors.

would allow to properly separate samples with low and high values of nitrate content, while the model developed following the second strategy would enable to obtain a good separation between low, medium and high values for this parameter. Likewise, according to Nicolaï et al. (2007), the models developed following strategies I and II would only discriminate between samples with low and high values of nitrate content. Significant differences ($P < 0.05$) were found between the SEP values obtained for both strategies (F and $F_{critical}$ data not shown), which highlights the importance of including part of the 2020 harvesting season samples' variability in the calibration set. In addition, it must be considered that one sample was removed from the V2 validation set since it showed a NH and T value over the established limits.

Although no significant differences ($P > 0.05$) were found when the SEP obtained using the MPLS and LOCAL algorithm were compared for both strategies (Table 5), it should be noted that for strategy II, the SEP decreased by 15% and the R^2_p increased by 29% when the non-linear regression algorithm was applied. This improvement in the results obtained when the LOCAL algorithm was used could be attributed to the distribution of the samples throughout the nitrate content range (Fig. 2B), which is not uniform, with the samples with nitrate content under 1500 mg/kg accounting for about 58% of the total population.

Ultimately, the influence of including different harvesting seasons, as well as the application of both a linear and a non-linear regression technique in the predictive capacity of the developed models, were assessed in this study and the outcomes could be interpreted as the best possible results achievable given the analytical methodology followed. This could be of great interest for the industry for the discrimination of plants which have a low, medium or high value of nitrate content and, therefore, to meet the current regulations (OJEU, 2011).

Although this methodology has been optimized by our group in previous studies, where we reported the great importance of minimising the time between the NIRS analysis and obtaining the reference data (Pérez-Marín et al., 2019), the reference method itself could be a limitation if we are looking for significant improvements in the prediction capacity of the models, since the reflectometer used to obtain the reference data could be considered as a semi-quantitative method for the estimation of nitrate content in the spinach plants. This issue could have a major impact on the prediction capacity of the models developed since, according to Fearn (1986), the prediction error in NIRS calibrations contains contributions from three sources, one being the error of the reference method, another the error of the NIR measurements and

Table 5
Comparison using Fisher's F test between the standard error of prediction values obtained with modified partial least squares –MPLS– and LOCAL regression algorithms for the two strategies tested.

Parameter	Strategy	Regression algorithm	SEP	F	$F_{critical}$
Nitrate content (mg/kg)	I	MPLS	920	1.04	1.41
		LOCAL	940		
	II	MPLS	889	1.40	1.52
		LOCAL	758		

SEP – Standard error of prediction.

finally the error of the model adjustment to describe the relationship between the reference and the NIR measurements.

In our view, according to these results and the previous results of our specific research topic for determining this safety parameter in spinach, the models developed here may be reaching the limits of precision and accuracy which can be achieved using this methodology, considering the low concentration of this parameter in spinach plants. Other approaches used as reference method to fit in with the NIR data, based on other more accurate analytical techniques at a quantitative level (e.g. high performance liquid chromatography, HPLC), could be explored, if our final objective is to quantify the nitrate content accurately. Nevertheless, from a practical point of view, it must be stressed that our results are of great use for the industry, allowing to classify the product on the production line, in real time, according to the current regulations.

4. Conclusions

The use of an online FT-NIR instrument along with a non-linear regression technique is a suitable alternative to predict the nitrate content in spinach plants in the industrial sorting lines, allowing to classify the product into the main categories or classes established by the official regulations, according to its commercial destination. A proper definition of the calibration set in terms of including the greatest variability along with the selection of the optimum chemometric treatment for a large data set –built with samples belonging to three different harvesting seasons showing a wide range of the parameter under consideration– is a paramount issue when developing quantitative NIR prediction models. Furthermore, the reference method used, the characteristic of the product and parameter to be analysed can be considered the major limitation when aiming at enhancing the models' prediction capacity.

CRediT authorship contribution statement

Miguel Vega-Castellote: Data acquisition, Methodology, Formal analysis, Investigation, Software, Data curation, Validation, Writing – original draft, Writing – review & editing, Visualization. **Dolores Pérez-Marín:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Irina Torres:** Data acquisition, Formal analysis, Investigation, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization. **María-Teresa Sánchez:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper in any way.

Acknowledgements

This research was under the research project 'Quality determination of spinach grown in Santaella (Córdoba)', funded by Gelagri Ibérica, S.L. The authors are grateful to Mrs. M[a] Carmen Fernández of the Animal Production Department for her technical assistance. Furthermore, the authors wish to express their gratitude to the Spanish Ministry of Universities for the support offered to Miguel Vega-Castellote in the form of the Training Programme for Academic Staff (FPU).

References

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near infrared diffuse reflectance spectra. *Applied Spectroscopy, 43*, 772–777. https://doi.org/10.1366/0003702894202201

Berzagui, P., Shenk, J. S., & Westerhaus, M. O. (2000). LOCAL prediction with near infrared multi-product databases. *Journal of Near Infrared Spectroscopy, 8*, 1–9. https://doi.org/10.1255/jnirs.258

Colla, G., Kim, H. J., Kyriacou, M. C., & Rouphael, Y. (2018). Nitrate in fruits and vegetables. *Scientia Horticulturae, 237*, 221–238. https://doi.org/10.1016/j.scienta.2018.04.016

Dambergs, R. G., Cozzolino, D., Cynkar, W. U., Janik, L., & Gishen, M. (2006). The determination of red grape quality parameters using the LOCAL algorithm. *Journal of Near Infrared Spectroscopy, 14*, 71–79. https://doi.org/10.1255/jnirs.593

Entrenas, J. A., Pérez-Marín, D., Torres, I., Garrido-Varo, A., & Sánchez, M. T. (2020). Simultaneous detection of quality and safety in spinach plants using a new generation of NIRS sensors. *Postharvest Biology and Technology, 160*, 1–8. https://doi.org/10.1016/j.postharvbio.2019.111026. Article 111026.

Fearn, T. (1986). Application of near infrared spectroscopy in the food industry. Some statistical comments on the errors in NIR calibrations. *Analytical Proceedings, 23*, 123–124.

Garrido-Varo, A., Sánchez-Bonilla, A., Maroto-Molina, F., Riccioli, C., & Pérez-Marín, D. (2018). Long-length fiber optic near-infrared (NIR) spectroscopy probes for on-line quality control of processed land animal proteins. *Applied Spectroscopy, 72*, 1170–1182. https://doi.org/10.1177/0003702817752111

Hruschka, W. R. (2001). Data analysis: Wavelength selection methods. In P. C. Williams, & K. H. Norris (Eds.), *Near-Infrared technology in the agricultural and food industries* (pp. 39–58) (American Association of Cereal Chemists, Inc., St. Paul, MN, USA).

Jaworska, G. (2005). Content of nitrates, nitrites, and oxalates in New Zealand spinach. *Food Chemistry, 89*, 235–242. https://doi.org/10.1016/j.foodchem.2004.02.030

Jerome, J., & Workman, J. (2008). NIR spectroscopy calibration basics. In D. A. Burns, & E. W. Ciurczac (Eds.), *Handbook of near infrared analysis* (pp. 123–150). New York-Basel, NY, USA: Marcel Dekker.

Mahanti, N. K., Chakraborty, S. K., Kotwaliwale, N., & Vishwakarma, A. K. (2020). Chemometric strategies for non-destructive and rapid assessment of nitrate content in harvested spinach using Vis-NIR spectroscopy. *Journal of Food Science, 85*, 3653–3662. https://doi.org/10.1111/1750-3841.15420

Naes, T., Isaksson, T., Fearn, T., & Davies, A. (2002). *A user-friendly guide to multivariate calibration and classification*. Chichester, UK: NIR Publications.

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., et al. (2007). Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology, 46*, 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Official Journal of the European Union (OJEU). (2011). *Commission regulation (EU) No. 1258/2011 of 2 December 2011 Amending regulation (EC) No 1881/2006 as regards maximum levels for nitrates in Foodstuffs*. OJ L 320/15-17, 3.12.2011.

Peirs, A., Tirry, J., Verlinden, B., Darius, P., & Nicolaï, B. (2003). Effect of biological variability on the robustness of NIR models for soluble solids content of apples. *Postharvest Biology and Technology, 28*, 269–280. https://doi.org/10.1016/S0925-5214(02)00196-5

Pérez-Marín, D., Fearn, T., Guerrero, J. E., & Garrido-Varo, A. (2012). Improving NIRS predictions of ingredient composition in compound feedingstuffs using Bayesian non-parametric calibrations. *Chemometrics and Intelligent Laboratory Systems, 110*, 108–112. https://doi.org/10.1016/j.chemolab.2011.10.007

Pérez-Marín, D., Garrido-Varo, A., & Guerrero, J. E. (2005). Implementation of LOCAL algorithm with near-infrared spectroscopy for compliance assurance in compound feedingstuffs. *Applied Spectroscopy, 59*, 69–77. https://doi.org/10.1366/0003702052940585

Pérez-Marín, D., Garrido-Varo, A., & Guerrero, J. E. (2007). Non-linear methods in NIRS quantitative analysis. *Talanta, 72*, 28–42. https://doi.org/10.1016/j.talanta.2006.10.036

Pérez-Marín, D., Garrido-Varo, A., Riccioli, C., & Fearn, T. (2018). *Training guidelines and wider exploitations of NIRS*. Food Integrity: Ensuring the Integrity of the European Food Chain, Deliverable: 19.7, Retrieved from https://secure.fera.defra.gov.uk/foodintegrity/secure/downloadFile.cfm?id=762 accessed . (Accessed 3 May 2021).

Pérez-Marín, D., Torres, I., Entrenas, J. A., Vega, M., & Sánchez, M. T. (2019). Pre-harvest screening on-vine of spinach quality and safety using NIRS technology. *Spectrochimica Acta: Molecular and Biomolecular Spectroscopy, 207*, 242–250. https://doi.org/10.1016/j.saa.2018.09.035

Proietti, S., Moscatello, S., Giacomelli, G. A., & Battistelli, A. (2013). Influence of the interaction between light intensity and CO_2 concentration on productivity and quality of spinach (*Spinacia oleracea* L.) grown in fully controlled environment. *Advances in Space Research, 52*, 1193–1200. https://doi.org/10.1016/j.asr.2013.06.005

Saeys, W., Do Trong, N. N., Van Beers, R., & Nicolaï, B. M. (2019). Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review. *Postharvest Biology and Technology, 158*, Article 110981. https://doi.org/10.1016/j.postharvbio.2019.110981, 1–19.

Sánchez, M. T., De la Haba, M. J., Guerrero, J. E., Garrido-Varo, A., & Pérez-Marín, D. (2011). Testing of a local approach for the prediction of quality parameters in intact nectarines using a portable NIRS instrument. *Postharvest Biology and Technology, 60*, 130–135. https://doi.org/10.1016/j.postharvbio.2010.12.006

Shenk, J. S., & Westerhaus, M. O. (1995a). *Analysis of agriculture and food products by near infrared reflectance spectroscopy*. USA: Monograph, NIRSystem, Inc., 12101 Road, Silver Spring. MD 20904.

Shenk, J. S., & Westerhaus, M. O. (1995b). *Routine operation, calibration, development and network system management manual*. USA: NIRSystems, Inc., 12101 Tech Road, Silver Spring. MD 20904.

Shenk, J. S., & Westerhaus, M. O. (1996). Calibration the ISI way. In A. M. C. Davies, & P. C. Williams (Eds.), *Near infrared spectroscopy: The future waves* (pp. 198–202). Chichester, UK: NIR Publications.

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy, 5*, 223–232. https://doi.org/10.1255/jnirs.115

Subedi, P. P., Walsh, K. B., & Hopkins, D. W. (2012). Assessment of titratable acidity in fruit using short wave near infrared spectroscopy. Part B: Intact fruit studies. *Journal of Near Infrared Spectroscopy, 20*, 459–463. https://doi.org/10.1255/jnirs.1011

Torres, I., Pérez-Marín, D., De la Haba, M. J., & Sánchez, M. T. (2017). Developing universal models for the prediction of physical quality in citrus fruits analysed on-tree using portable NIRS sensors. *Biosystems Engineering, 153*, 140–148. https://doi.org/10.1016/j.biosystemseng.2016.11.007

Torres, I., Sánchez, M. T., De la Haba, M. J., & Pérez-Marín, D. (2019). LOCAL regression applied to a citrus multispecies library to assess chemical quality parameters using near infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 217*, 206–214. https://doi.org/10.1016/j.saa.2019.03.090

Torres, I., Sánchez, M. T., Entrenas, J. A., Garrido-Varo, A., & Pérez-Marín, D. (2020a). Monitoring quality and safety assessment of summer squashes along the food supply chain using near infrared sensors. *Postharvest Biology and Technology, 154*, 21–30. https://doi.org/10.1016/j.postharvbio.2019.04.015

Torres, I., Sánchez, M. T., & Pérez-Marín, D. (2020b). Integrated soluble solid and nitrate content assessment of spinach plants using portable NIRS sensors along the supply chain. *Postharvest Biology and Technology, 168*, 1–7. https://doi.org/10.1016/j.postharvbio.2020.111273. Article 111273.

Torres, I., Sánchez, M. T., Vega-Castellote, M., Luqui-Muñoz, N., & Pérez-Marín, D. (2021). Routine NIRS analysis methodology to predict quality and safety indexes in spinach plants during their growing season in the field. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 246*, 1–7. https://doi.org/10.1016/j.saa.2020.118972. Article 118972.

Vega-Castellote, M., Pérez-Marín, D., Torres, I., Moreno-Rojas, J. M., & Sánchez, M. T. (2021). Exploring the potential of NIRS technology for the *in situ* prediction of amygdalin content and classification by bitterness of in-shell and shelled intact almonds. *Journal of Food Engineering, 294*, 1–10. https://doi.org/10.1016/j.jfoodeng.2020.110406. Article 110406.

Walsh, K. B., McGlone, V. A., & Han, D. H. (2020). The uses of near infra-red spectroscopy in postharvest decision support: A review. *Postharvest Biology and Technology, 163*, 1–12. https://doi.org/10.1016/j.postharvbio.2020.111139. Article 111139.

Williams, P. C. (2001). Implementation of near-infrared technology. In P. C. Williams, & K. H. Norris (Eds.), *Near-infrared technology in the agricultural and food industries* (pp. 145–169). St. Paul, MN, USA: AACC Inc.

Windham, W. R., Mertens, D. R., & Barton, F. E., II (1989). Protocol of NIRS calibration: Sample selection and equation development and validation. In G. C. Martens, J. S. Shenk, & F. E. Barton, II (Eds.), *Near infrared spectroscopy (NIRS): Analysis of forage quality. Agriculture handbook nº643* (pp. 96–103). Washington, DC, USA: USDA-ARS, Government Printing Office.