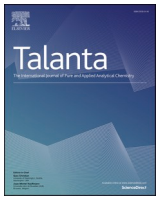
Talanta 276 (2024) 126271

Contents lists available at ScienceDirect

Talanta

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



Measurement errors and implications for preprocessing in miniaturised near-infrared spectrometers: Classification of sweet and bitter almonds as a case of study

Jokin Ezenarro [a], Jordi Riu [a], Hawbeer Jamal Ahmed [a,b], Olga Busto [a], Barbara Giussani [c,**], Ricard Boqué [a,*]

[a] Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, 43007, Tarragona, Catalonia, Spain
[b] United Science Colleges, Department of Chemistry, Bakhan 108, Sulaymaneyah, Iraq
[c] Dipartimento di Scienza e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio, 9, 22100, Como, Italy

ARTICLE INFO

Handling Editor: A Campiglia

Keywords:
Preprocessing
Error covariance matrices
Correlation error
Variability sources
Near-infrared (NIR)
Discriminant analysis

ABSTRACT

Near-infrared (NIR) spectroscopy is a well-established analytical technique that has been used in many applications over the years. Due to the advancements in the semiconductor industry, NIR instruments have evolved from benchtop instruments to miniaturised portable devices. The miniaturised NIR instruments have gained more interest in recent years because of the fast and robust measurements they provide with almost no sample pretreatments.

However, due to the very different configurations and characteristics of these instruments, they need a dedicated optimization of the measurement conditions, which is crucial for obtaining reliable results. To comprehensively grasp the capabilities and potentials offered by these sensors, it is imperative to examine errors that can affect the raw data, which is a facet frequently overlooked. In this study, measurement error covariance and correlation matrices were calculated and then visually inspected to gain insight into the error structures associated with the devices, and to find the optimal preprocessing technique that may result in the improvement of the models built.

This strategy was applied to the classification of sweet and bitter almonds, which were measured with the three portable low-cost NIR devices (SCiO, FlameNIR+ and NeoSpectra Micro Development Kit) after removing the shelled, since their classification is of utmost importance for the almond industry. The results showed that bitter almonds can be classified from sweet almonds using any of the instruments after selecting the optimal preprocessing, obtained through inspection of covariance and correlation matrices. Measurements obtained with FlameNIR + device provided the best classification models with an accuracy of 98 %. The chosen strategy provides new insight into the performance characterization of the fast-growing miniaturised NIR instruments.

1. Introduction

Near-infrared (NIR) spectroscopy is a type of vibrational spectroscopy that detects changes in the vibrations of molecules in response to electromagnetic radiation. This non-invasive method harnesses the interaction of near-infrared light with matter to reveal valuable information about the composition and molecular structure of various substances. As almost all the organic molecules absorb light in the mid-infrared region and this is reflected in the overtone bands that appear in the near-infrared region, this technology is especially valuable when analysing food samples, which are mostly made of organic matter [1,2]. Furthermore, vibrational spectroscopy has some advantages over other analytical techniques because it can monitor analytes with high selectivity without the need for time-consuming and usually not-green sample treatments [3,4]. Because it is robust and non-destructive, NIR spectroscopy has become a successful analysis tool in many fields, such as the agri-food, pharmaceutical, and polymer industry [5].

One of the main challenges in using NIR techniques is that a

* Corresponding author.
** Corresponding author.
E-mail addresses: barbara.giussani@uninsubria.it (B. Giussani), ricard.boque@urv.cat (R. Boqué).

https://doi.org/10.1016/j.talanta.2024.126271
Received 18 December 2023; Received in revised form 14 May 2024; Accepted 15 May 2024
Available online 16 May 2024

spectrum consists of a number of bands emerging from overtones and combination modes that substantially overlap with each other, making it difficult to analyse the information contained in them. Another major challenge in NIR spectroscopy is the presence of baseline shifts and drifts, which needs to be corrected in order to build proper models. Consequently, chemometric tools are commonly employed during data analysis to correct and extract information from NIR spectra. These techniques help overcome the mentioned problems caused by the high complexity of the raw signal, which is of multivariate nature [6,7].

The use of chemometrics in spectroscopy is growing as the recent advancements in the semiconductor industry have led to the miniaturisation of classical laboratory instruments into handheld spectroscopic devices. The downsizing of the spectrometers is, among others, due to incorporating novel technological solutions, e.g., based on MEMS (micro-electro-mechanical systems) and MOEMS (micro-opto-electro-mechanical systems). The fabricated miniaturised spectrometers have lower costs compared to benchtop instruments, which allows their use by a broader range of users and applications. For example, rapid analysis of milk [8], classification of edible oils [9] or prediction of nutritional parameters in insect powders [10]. These spectrometers follow the objectives of green analytical chemistry, providing rapid, non-destructive, and on-field analysis with almost no sample pretreatments and minimal use of reagents [11]. These on-field analyses with miniaturised spectrometers are of utmost importance in the industry because using conventional benchtop instruments requires transferring the sample to a laboratory, where the sample may undergo alteration and the obtention of the results may be delayed. The real-time monitoring in the production facilities ensures quality and safety matters and allows rapid intervention when a problem is detected. Moreover, it allows fast quality control checks by regulatory bodies in the markets and factories [12–14].

The aforementioned reduction in the dimension and cost of the spectrometers may result in lower performances compared with bulky instruments. The benchtop NIR instruments are mature devices that have been well studied and characterized in the past 20 years. They have uniform performances, rely on the same types of instrumentation such as the light source, and the measurements are made in the controlled conditions of the laboratory. The spectral wavelengths of the NIR region range from 800 to 2500 nm (12,500 to 4000 cm^{-1}), and benchtop devices usually cover the entire region. On the other hand, portable NIR spectrometers cover a portion of the NIR region and have lower spectral resolutions. They rely on different and new solutions due to the engineering difficulties of miniaturisation aspects. Such diverse technologies cause non-uniform performance and require more device-specific optimizations [5,15,16].

There has been an increasing effort to stablish the performance parameters of these miniaturised NIR devices in recent years. Some of the articles found in the literature focus on the assessment of the classification, identification or predictive abilities of these devices applied to different fields and their performance comparison with benchtop devices [17–23]. However, since the field of miniaturised NIR spectrometers is very new and evolving very fast, the sources of variability associated to these instruments are not yet well known for the different applications. So, a complete characterization of these newly developed spectrometers still needs to be explored and adapted to different types of samples, although there are some attempts to transfer the calibration data from benchtop instruments to miniaturised devices [24,25]. There are a few studies that have investigated the sources of variability in miniaturised devices that affect their performance, to understand the underlying error structures and develop optimal strategies for new measurements and applications [26–30]. Thus, these miniaturised spectrometers require a more in-depth evaluation of the measurement errors, caused by the various sources of variability present in the data acquisition, to get the best performance characterization that can produce reliable models.

The main objective of this work was to characterize the errors associated with three low-cost miniaturised NIR instruments, of different robustness and price, when analysing whole almonds: SCiO (Consumer Physics, few thousand euros + software), FlameNIR+ (Ocean Optics, around ten thousand euros), and NeoSpectra Micro Development Kit (Si-Ware Systems, a couple thousand euros + periodic subscription). Each of the three NIR spectrometers uses different technological solutions and covers a different region of the NIR spectra so that the data are complementary and just overlap in a very narrow part of the spectra. The study identifies underlying error types and structures through the analysis of multivariate measurement errors, which is a well-established statistical way to characterize error structures by building error covariance and correlation matrices from replicas of the spectra. The effectiveness of multivariate measurement error and its incorporation during data analysis for other analytical instruments has been studied, but its application to miniaturised NIR spectrometers is very limited and missing in the literature [29,31–33]. Recently, Gorla et al. [28] tried to reveal the error sources in one miniaturised instrument and used this information to determine different properties of forage samples. Similarly, Wentzell et al. [29] studied the error structures present in NIR spectra of wood samples for differentiating the tree species. They concluded that by evaluating error structures present in the data, optimal analytical strategies can be developed.

As a second objective, the best spectral preprocessing for each instrument was identified through the inspection of error covariance and correlation plots. Finally, to test the effectiveness of the proposed method, the performances of all three instruments were evaluated to classify bitter and sweet almonds, which has already been tried with not so low-cost miniaturised NIR spectrometers [34,35]. One of the most important aspects of the almond industry is the discrimination of bitter almonds from sweet ones since it affects their commercialization and usage in a variety of foods. Apart from the unpleasant taste, bitter almonds have serious health risks because they contain toxic compounds such as amygdalin, whose hydrolysis creates benzaldehyde and cyanide, the latter causing poisoning and potentially accidental death.

2. Materials and methods

2.1. Instrumentation

All measurements were performed using three portable NIR spectrometers: SCiO (Consumer Physics, Herzliya, Israel), FlameNIR+ (Ocean Optics, Dunedin, USA), and NeoSpectra Micro Development Kit (Si-Ware Systems, Cairo, Egypt). The working principle and the instrumental solutions are different in the three devices.

SCiO is a pocket-size NIR spectrometer that has a weight of 35 g and its dimensions are 67.7 × 40.2 × 18.8 mm. It acquires spectra in the 740–1070 nm wavelength range with interpolated spectra with a resolution of 1 nm (331 variables) [36,37]. It can perform measurements both in contact and distance mode and is usually used for solid samples. The distance between the sample and the device should be less than 10 mm. SCiO should be connected to a smartphone via Bluetooth, and the spectra are recorded using 'The Lab' app available for Android and iOS systems. The app does not allow setting any measurement parameters, and the scan time is less than 5 s. The spectra can then be downloaded from the private area of a cloud web address (thelab.consumerphysics.com). The device should be calibrated each time it is turned on using a calibration standard on the back cover of the device. The device can be used in two operation modes: connected to a power supply or running on battery. Regarding technological solutions, SCiO has a light source based on light-emitting diodes (LEDs), making the device more cost-effective and decreasing power consumption. SCiO contains a silicon detector based on complementary metal-oxide-semiconductor (CMOS) in the form of a 4 × 3 photodiode array with optical filters over the individual pixels. No initial warm-up is required. To perform the measurements, the SCiO device was fully charged and background acquisitions were acquired before starting each measurement session.

The spectra were acquired in reflectance mode by directly pointing the device onto the almonds at a fixed distance of around 0.5 cm, fixed geometry and avoiding the tilting (Fig. 1a).

The FlameNIR+ is a miniaturised NIR spectrophotometer that weights 989 g (including all modules) and its dimensions are 89.1 × 63.3 × 31.9 mm. The captured spectra cover a range between 970 and 1700 nm with a resolution of 6 nm (128 variables). As it is a modular spectrometer consisting of different probes and optic fibres, several measurement modes can be arranged: surface measurement in direct or diffuse reflectance, or transmittance/absorbance or fluorescence measurements in a cuvette. The spectrometer is powered and controlled by a computer connected via a USB cable; instead, the NIR light source is directly connected to a power supply. The number of scans, integration time and measurement modes can be changed in the freely available 'OceanView' software. For an adequate measurement, a black reference and a white reference must be collected first, which, in the case of diffuse reflectance measuring arrangement, consists of measuring a spectrum with the NIR light source turned off and another spectrum of the Spectralon™ (Labsphere, Sutton, USA) sample with the source turned on. As suggested by the manufacturer, the white or blank reference must be periodically repeated during the usage as the temperature and measurement conditions may vary. To reduce this variation, a previous warm-up of the NIR light source is recommended. The detector of the spectrometer consists of an uncooled InGaAs array. The measurements with the FlameNIR + device were performed in reflectance mode, using an optic fibre probe located in a probe holder and attached to the light source and the spectrometer (Fig. 1b).

The NeoSpectra device has a weight of 17 g and dimensions of 32 × 32 × 22 mm. The wavelength range is from 1350 to 2558 nm (134 variables) with a varying spectral resolution that ranges from 5 nm to 17 nm as wavelength numbers increase. Only contact measurements can be performed, as optimal signal is acquired when the sample is in contact with the window containing the light source and sensor. The device must be connected to a computer via a USB cable, and the spectra are collected by a software (SpectroMOST Micro) and are stored in the computer. It allows configuring some parameters such as the scan time, run mode (single or continuous measurements), display mode (reflectance or absorbance), and data interpolation in each spectrum collected. A reflection standard such as Spectralon™ is required when the software is started or when any operational parameter is changed. The instrumental design of NeoSpectra consists of a light source made of three halogen tungsten lamps that require an initial warm-up before performing the measurements to stabilize the light intensity. The wavelength selector is a Michelson interferometer made by the MEMS technique and has a single InGaAs photodetector [21,22]. For the NeoSpectra measurements, background acquisitions were performed at the beginning of the measurements and every hour thereafter, because NeoSpectra needs frequent background resets due to the heating of optical components [28]. An initial warm-up of 20 min, simply using a continuous measurement setting but without placing any sample on the window, was performed before each measurement session, then, the scanning time was set to 2 s for the samples. The almonds were put directly on the NeoSpectra window (Fig. 1c) and were acquired in reflectance mode.

All measurements with the three devices were performed at room temperature under ambient light. The average raw spectra of the sweet and bitter almonds obtained with the three devices are shown in Fig. 2. Additionally, the standard deviation of the spectra of each device can be seen in supplementary Fig. S1.

2.2. Samples

The samples included different varieties of almonds from La Palma d'Ebre, in Tarragona, Catalonia, Spain. All almonds were from the same harvesting season and were collected in September 2022. A total of 150 almonds were analysed after removing the shell, of which 75 were sweet almonds, and 75 were bitter almonds. The sweet almonds were from three different varieties; 25 were of the *Ferragnes* variety (group S1) and 50 were from two local varieties of *Comuna* almonds (groups S2 and S3), which are the most processed in the Spanish industry [38]. The bitter almonds belonged to three different non-identified varieties, with different external morphology, coming from different fields of the same region (groups B1, B2 and B3).

For a preliminary study of error sources, 18 almonds were randomly chosen: 9 sweet and 9 bitter, 3 of each variety. Two different types of replicates were used in the measurements: replacement replicates, that is, changing the position of the sample each time (randomly, assuring both faces of the almond are positioned at least once) when it is measured, to account for variations related to the sample position; and instrumental replicates, that is, measuring the spectra without moving the sample to account for variations related to the instrument [29]. Three replacement replicates and five instrumental replicates for each position were measured per each sample, in the same analytical session, accounting for the heterogeneity of the samples and instrumental variations.

Upon analysing the outcomes of the initial experiment, the rest of

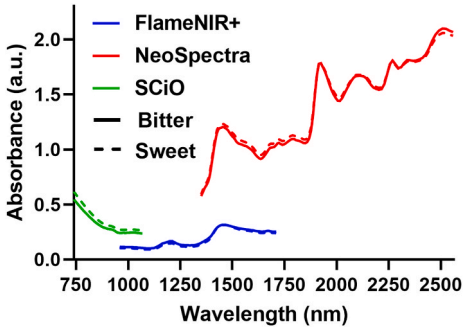


Fig. 2. Average raw spectra of sweet and bitter shelled almonds measured with SCiO, FlameNIR+ and NeoSpectra.

Fig. 1. Almond measurement set-ups for a) SCiO, b) FlameNIR+ and c) NeoSpectra devices.

almonds were measured randomly along 10 days, to include the variability of the measurement session into the models as happens in real-working conditions. Three repositioning replicates were acquired randomly, ensuring that each face was analysed at least once. The spectra of the replacement replicates of the almonds used for the preliminary study (three of each almond) were added to this data set. The spectra of all replicates for each almond were further used for data analysis and modelling by assigning them the class of the almond they belong to.

2.3. Statistical data analysis

MATLAB R2021b (Mathworks Inc., Natick, MA, USA) and PLS_Toolbox version 9.0 for MATLAB (Eigenvector Inc, Manson, WA, USA) installed on a PC with Windows operating system were used for data analysis. Data were organized in an **X** matrix for each instrument, containing the samples in the rows and the wavelengths in the columns. In the case of SCiO and NeoSpectra, the spectra were transformed from reflectance to absorbance units. In the case of FlameNIR+, the sensor directly provided absorbance values. Error Covariance Matrices (ECMs) were calculated using in-house routines with MATLAB 2021b, while Partial Least Squares Discriminant Analysis (PLS-DA) was performed with the PLS_Toolbox 9.0. For the PLS-DA models, a **Y** vector was defined, containing dummy values for the classes (zeros for sweet almonds and ones for bitter almonds). Different spectral pre-processing methods were tested but only the most relevant are shown in this article: detrending, standard normal variate (SNV), and first and second Savitzky–Golay derivatives (2nd order polynomial, 15-point window; after optimization) and the combination of these two with SNV. After spectral pre-processing, data were finally mean-centred in all calculations. The original data set was split into a calibration set and a test set using the onion algorithm, as implemented in the PLS_Toolbox: 2/3 of the samples were maintained in the calibration set making sure all types of almonds were well and equally represented in both sets. To validate the multivariate models, both cross-validation (5 data splits and 5 iterations – on calibration set) and external validation were used. The selection of the best models was determined by striking a balance between minimizing the number of latent variables (LVs), and maximising both sensitivity (samples belonging to a class correctly assigned to that class) and specificity (samples not belonging to the class correctly not assigned to that class) all of which collectively contribute to accuracy (correctly assigned samples).

2.3.1. Error covariance matrix (ECM)

ECMs are a practical way to characterize multivariate measurement errors by describing the relationships between measurement errors across the channels/wavelengths [26]. An ECM is a symmetric matrix, in which the diagonal elements contain the variance of the measurement error at each channel and the off-diagonal elements contain the covariance of the errors between pairs of channels. ECMs are typically represented graphically, and their visual analysis provides interesting information on the magnitude and type of errors. The ECM is a useful tool to find the structure associated to the measured errors (e.g., proportional errors, constant errors, …) so that the choice of the optimal data preprocessing may be derived. When there is insufficient prior knowledge about the error types and structures in a measurement (as is the case in this work with miniaturised instruments), then the ECMs are calculated from the experimental estimation method, which is based on the analysis of the replicates. To calculate the covariance matrix using this method, the (approximate) true spectrum of a sample is estimated from the mean of the spectral replicates of that sample. Then, a residual matrix is calculated by subtracting the estimation of the real value from each replicate spectrum. Finally, the error covariance matrix is calculated as the covariance between the residuals, as shown in Eq. (1):

$$\Sigma_{cov} = \frac{1}{(n-1)} \sum_{k=1}^{n} (\mathbf{X}_k - \overline{\mathbf{X}})^{\mathrm{T}} (\mathbf{X}_k - \overline{\mathbf{X}}) \quad (1)$$

where $\mathbf{\Sigma}_{\mathrm{cov}}$ is the covariance matrix of the ith sample, n is the number of replicates of the ith sample, $\mathbf{X}_k$ is the measured spectrum of the kth replicate for the ith sample, and $\overline{X}$ is the mean spectrum of n replicates.

The covariance matrix depends on the magnitudes of the variances, and this makes their visual interpretation difficult when there are few channels with significantly higher variances (e.g., certain wavelengths with much lower precision than the rest of wavelengths). Variations among other channels can be hidden. For this reason, error correlation matrices were also calculated by scaling the variances and thus removing the effects of magnitudes of variations. The error correlation matrix, with all the values scaled between +1 and −1, contains the correlation coefficients of the covariance matrix and is calculated as shown in Eq. (2).

$$\Sigma_{\mathrm{corr}} = \Sigma_{\mathrm{cov}} \cdot \Big/ \sqrt{diag(\Sigma_{\mathrm{cov}}) diag(\Sigma_{\mathrm{cov}})^{\mathrm{T}}} \quad (2)$$

Unless the number of replicates for a sample is high (>20 if independent and identically distributed errors are assumed), the error covariance matrix has a certain degree of uncertainty. For this reason, it is important to have a sufficient number of replicates or otherwise to pool the error covariance over different subsets of samples by taking the mean of all covariance matrices ($\mathbf{\Sigma_{pooled}}$). The pooling solution is generally preferred with NIR spectra because the measurement data do not change very much for the same types of samples [33]. In this particular case, covariance matrices were pooled over the 150 almonds.

2.3.2. Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA is a supervised method used to classify samples based on specific properties assigned as different classes. In PLS-DA, a PLS regression model that links the independent variables (**X** matrix of NIR spectra in our case) to a vector **Y** containing the assigned classes as integer numbers is calculated. In this case, 0 was assigned to indicate sweet almonds, and 1 was assigned for bitter almonds. An unknown sample is classified using the projected value of the PLS model. This value, which is a real number rather than an integer, ought to ideally be near to the values used to define the class (here either 0 or 1). A cut-off value or a threshold, between 0 and 1, is established to maximize the selectivity and specificity of the model; so that an unknown sample is assigned to class 1 if the prediction is larger than the cut-off value, or assigned to class 0 if it is lower than the cut-off value. This threshold is set by minimizing the probability of both false positives and false negatives (assuming that the predicted values for each class are approximately normally distributed), using the algorithm implemented in the PLS_Toolbox. Additionally, for the construction of the models, the right number of latent variables (LVs) must be chosen to prevent underfitting or overfitting the models. The LVs are linear combinations of the initially selected variables that maximize the discrimination among the groups [39]. In this work, the number of LVs was chosen considering a compromise between the complexity of the model and the accuracy in prediction, in order to avoid overfitted models.

3. Results and discussion

3.1. Preliminary study of error sources

In the study of error covariance and correlation matrices, it is crucial to identify the type of replicates that will have an impact on these matrices. Instrumental variations are expected to have a smaller effect on the measurement errors than the sample variations due to the heterogeneity of the sample, as seen in previous works [28]. To corroborate this, for all three instruments, error covariance matrices were calculated with the spectra acquired from the 18 almonds selected for this

preliminary study, shown in Fig. 3. Additionally, error correlation matrices for the same spectra can be seen in supplementary Fig. S2.

As Fig. 3 shows, spectra of both types of almonds recorded with all three spectrometers gave similar error structure. The effects that might influence the measurements affect in a comparable scale the spectra of both types of almonds. Spectra of sweet and bitter almonds can thus be used together in the estimation of error covariance and correlation matrices, to find an optimal preprocessing method that will work for all considered almonds.

Furthermore, from Fig. 3 it can be concluded that the effects observed in the error covariance matrices of instrumental replicates are negligible when compared to the replacement replicates, as the latter ones have a considerably bigger scale (around 3 orders of magnitude, with the exception of the central point in NeoSpectra, related to sample heating). This is an expected result, given that the measurements obtained from an instrument, including a low-cost miniaturised instrument, exhibit minimal fluctuations during the timeframe in which these spectra were recorded. However, it is important to note that when performing the measurement, the exact positioning, e.g., angle and distance between the sensor and the sample, may not be consistently replicated and this may be reflected in the error structures. Based on these results, the almonds were analysed in triplicate, performing only repositioning replications.

3.2. Multivariate statistical analysis

3.2.1. Error covariance matrix (ECM)

The error covariance and correlation plots, calculated for the replacement replicates of each sample and averaged for all 150 samples, were visually inspected to understand the error types and structures for the three sensors. Fig. 4 shows the error covariance and error correlation matrices of the three instruments.

In the case of SCiO, the errors seem to be somehow homoscedastic, which is a type of error that has a uniform variance across the channels on the diagonal of the covariance matrix. This can be seen from Fig. 4a, in which diagonal elements have similar values. In addition, the errors were also highly correlated (Fig. 4d), which means that there is a relationship among the errors for different variables since most of the correlation matrix was close to one. A constant offset noise was observed, which is a type of correlated noise that can be seen in all three instruments and it is usually caused by temperature drifts or changing light scattering effects [40]. These effects shift the entire signal by a constant value and can be seen from the fact that covariance values were non-zero across all channels. This offset noise might be due to the repositioning of the almonds between replacement replicate scans. Additionally, a multiplicative noise can also be observed from the fact that the error covariance is proportional to the spectral signal (e.g., the peaks between 900 and 1000 nm) by comparing covariance matrix from Fig. 4a with the spectral signal from Fig. 2. Constant offset and multiplicative noise are typical characteristics of NIR spectra, which can be caused by different variations in the sample or the instrument [28,40].

Regarding the measurements with FlameNIR+, the magnitude of errors in the covariance matrix was smaller compared to the other devices, suggesting better measurement stability. On contrast, the errors seem to be heteroscedastic, this is, inconsistent across the diagonal of the matrix (Fig. 4b). The error types were similar to the other instruments, with a high correlation among them (Fig. 4e), additionally, a constant offset was observed; especially in the 1440–1600 nm region, which later can be corrected with appropriate preprocessing.

As for NeoSpectra measurements, the error structures showed more heteroscedastic errors across all regions (Fig. 4c). The noise proportional to the spectral signal was also present. Errors were higher in magnitude than in FlameNIR+, although a bit lower than in SCiO. This was already predicted from the noisier spectra given by NeoSpectra, especially in the range of 1400–1800 nm (supplementary Fig. S1Supplementary Fig. S1). The higher errors, such as the peak present around 1900–1950 nm, which corresponds to the first overtone of the water absorption band, is probably related to the heating up of the sample and its dehydration during the measurement, as this instrument not only enlightens the measured point but a broad area (see Fig. 1c) [41].

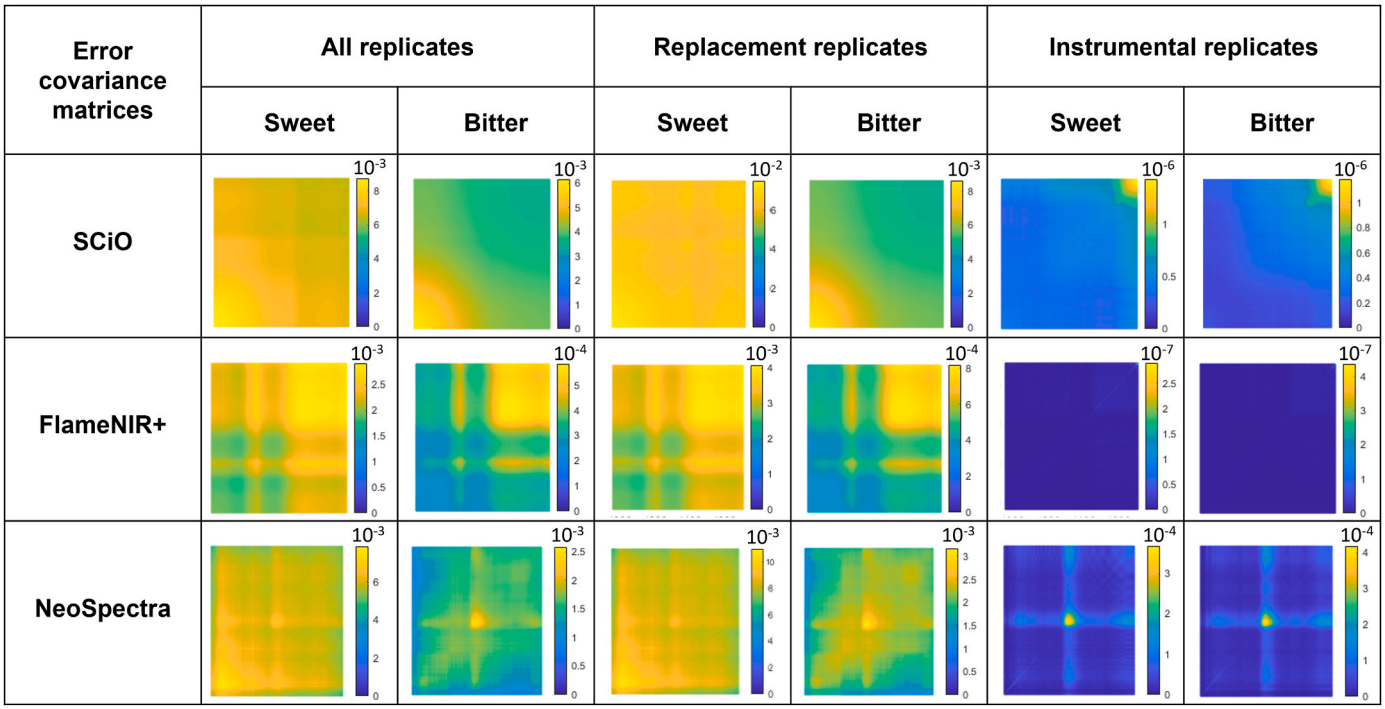


Fig. 3. Error covariance matrices for the SCiO, FlameNIR+ and NeoSpectra spectrophotometers considering different types of replicates: replacement, instrumental and both; averaging the matrices for sweet and bitter almonds separately.

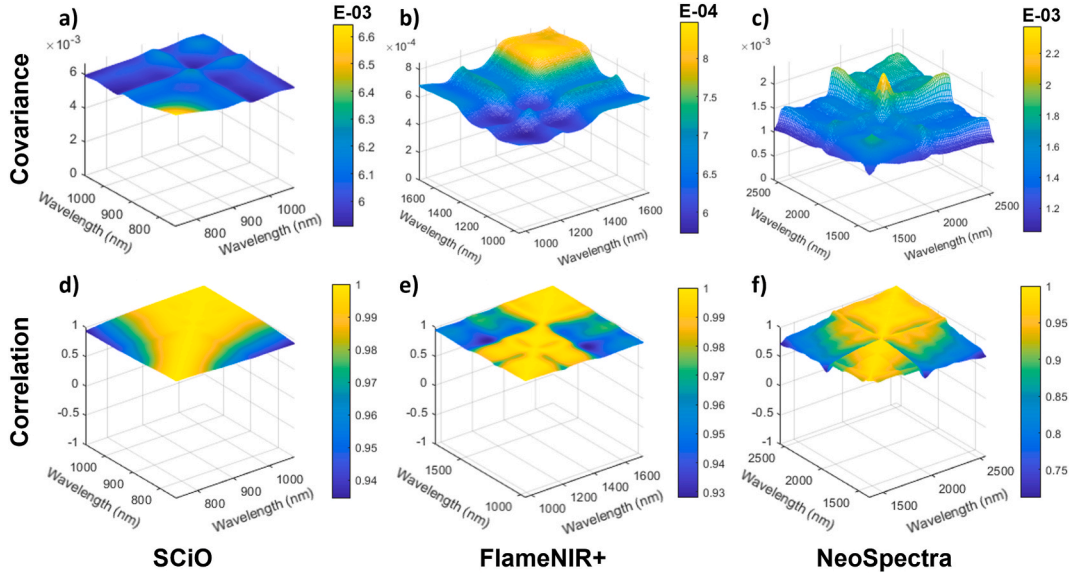


Fig. 4. Measurement error covariance and correlation matrices calculated from spectra acquired with SCiO (a,d), FlameNIR+ (b,e) and NeoSpectra (c,f), respectively. The matrices are pooled over the whole set of 150 samples, both sweet and bitter with the replacement replicates.

3.2.2. *Spectral preprocessing*

Different preprocessing methods were applied to the whole data set of 150 almonds, while evaluating the plots to see their effect on the noise structures. For this, it is hypothesized that the preprocessing method that is more capable of removing the heteroscedasticity in the data will result in the best performing classification model, as a normal distribution of the errors is one of the assumptions of the PLS model calculation [28]. In an attempt to correct the error structures present in the data, the most common preprocessing methods used in infrared spectroscopy were tested [40].

Fig. 5 shows the error correlation matrices for different preprocessing methods. In general, it can be observed that preprocessing the data results in smoother surfaces and less correlated (more randomly distributed) errors, with the higher correlations only being present in the diagonal of the matrix. At the same time, preprocessing makes the error more homoscedastic, as the variance of the errors becomes more equally distributed through all the spectrum (supplementary Fig. S3). Derivatives and derivatives combined with SNV seem to be the best selection, as the plots show that the errors have a more random distribution compared to raw data or other preprocessings. From this it is postulated that these preprocessing methods would be the most adequate to apply for building the classification models of almonds.

3.2.3. *Classification of almonds*

PLS-DA models were built for each set of spectra with a different preprocessing method and their performance was evaluated using the test set. For building the models for the FlameNIR+ and NeoSpectra instruments, the last part of the spectra (>1650 nm and >2400 nm, respectively) was removed, as a better performance was obtained. In addition, a cross-validation was carried out in the calibration set using ten iterations of the 5-fold random subset validation. Table 1 shows the performance indicators of the external validation set for each

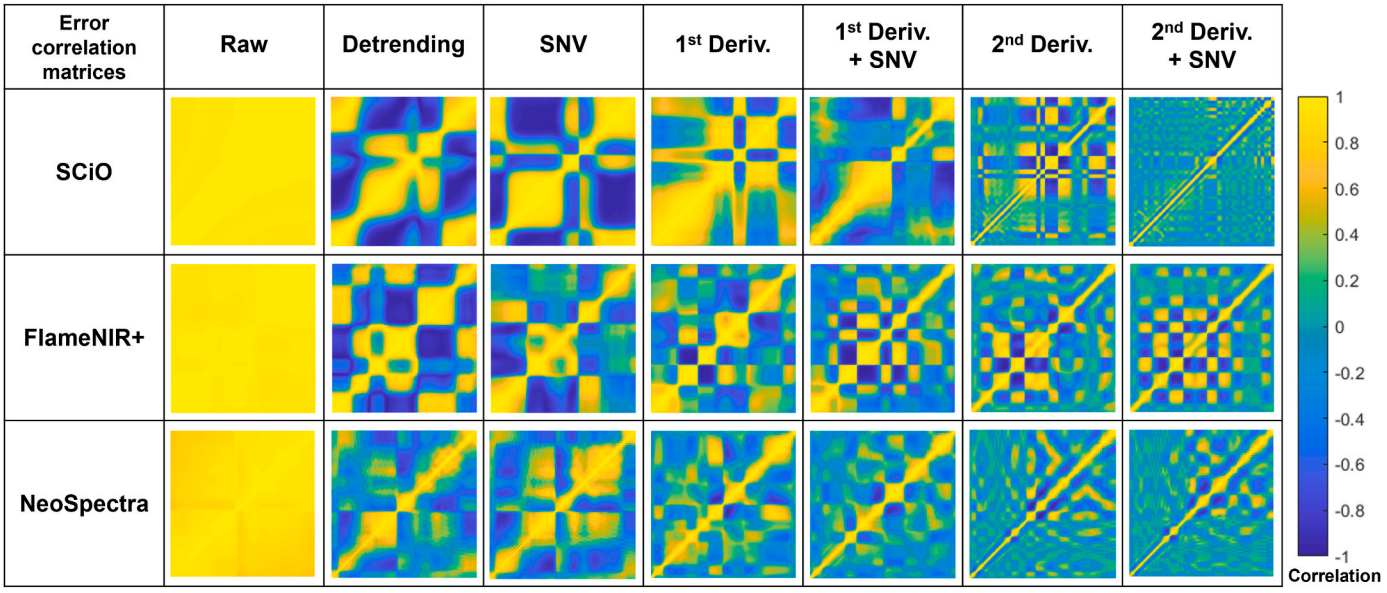


Fig. 5. Error correlation matrices calculated from raw spectra and using different preprocessing methods for different spectrophotometers, where yellow colour shows the highest positive correlation.

Table 1
Performance indicators of the PLS-DA models built for each spectrophotometer using different preprocessing methods, where the model selected as optimal for each instrument is marked in bold and italics. #LV: number of Latent Variables; TPR: True Positive Rate (correctly classified bitter almonds); TNR: True Negative Rate (correctly classified sweet almonds); ACC: Classification Accuracy (total number of correctly classified almonds).

	SCiO				FlameNIR+				NeoSpectra			
	#LV	TPR	TNR	ACC	#LV	TPR	TNR	ACC	#LV	TPR	TNR	ACC
Raw	3	0.86	0.48	0.69	3	0.72	0.88	0.80	3	0.87	0.80	0.84
Detrend	3	0.86	0.44	0.67	3	0.68	0.88	0.78	2	0.87	0.85	0.86
SNV	4	0.71	0.65	0.69	4	0.80	0.92	0.86	2	0.87	0.85	0.86
1st Deriv.	3	0.82	0.70	0.76	4	0.76	1.00	0.88	2	0.90	0.85	0.88
1st Deriv.+SNV	4	0.88	0.76	0.82	4	0.92	0.96	0.94	3	0.90	0.85	0.88
2nd Deriv.	3	0.96	0.88	0.92	***4***	***0.96***	***1.00***	***0.98***	2	0.97	0.80	0.90
2nd Deriv.+SNV	***4***	***0.96***	***0.92***	***0.94***	4	0.92	0.96	0.94	***2***	***1.00***	***0.80***	***0.92***

classification model.

As it can be seen, the accuracy of the model is related with the homoscedasticity and randomness of the error present in the spectral replicates used for modelling. The greater the degree of diagonalisation in the error correlation matrix, the greater the accuracy achieved by PLS-DA models, when using the same spectrometer and a similar number of latent variables (LVs). This indicates that the character of the errors (heteroscedasticity or homoscedasticity; and randomness or correlation) of the input data in fact does affect the outcome of the model, and that the analysis of error correlation matrices can serve as a valuable approach to assess the suitability of a preprocessing method before constructing prediction or classification models [42].

However, the error correlation matrices must be considered as an indicative tool and not as a rule to follow, as there are more factors involved in providing a successful classification or prediction model. Although this study focused on plotting these matrices and the error distribution, complementary techniques have been proposed to help the analyst better describe or use these matrices, such as error distribution histograms [30]. In the end, the implications of using different preprocessing methods should be further studied by an analyst and the final models should be properly validated.

For this particular case, from Table 1 can be deduced that the best results were obtained for the FlameNIR + instrument, which offered a model with a global classification accuracy of 98 %. However, very good results can also be achieved with the other two, much cheaper sensors. From this it can be concluded that it is possible to use a portable NIR instrument for classifying bitter and sweet almonds with a high performance. Furthermore, this is accomplished with a rather simple data preprocessing step and a parsimonious model. In addition, it can be noted that the performance of the models is not affected by the fact of having different varieties of almonds, as they are capable of modelling all the groups adequately. Fig. 6 shows the discrimination plot of this classification model for the FlameNIR + device, that is, preprocessed using the second derivative (Savitzky-Golay, 2nd order polynomial, 15-point window).

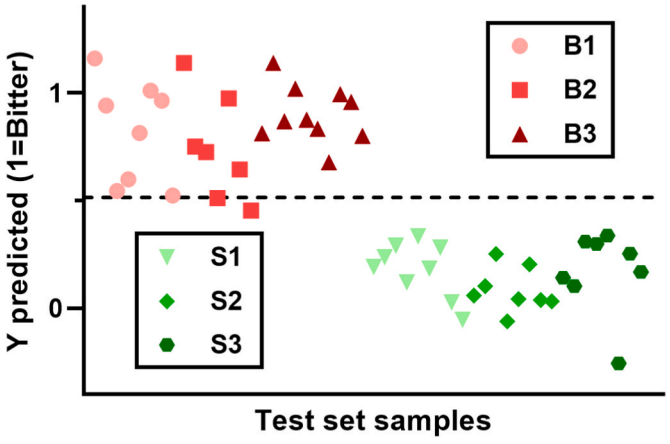


Fig. 6. PLS-DA discrimination plot for the model with highest accuracy for the prediction of the different varieties of bitter and sweet almonds (coloured in groups with different symbols) present in the test set, measured with FlameNIR+. The threshold is marked with a dotted line.

4. Conclusions

Error covariance and correlation matrices have been proven to be a method with high potential to study and quantify the errors present in the data provided by miniaturised instruments, offering a structured characterisation of the sources of variability that may influence the performance of three low-cost miniaturised NIR spectrometers, in this case.

The visualisation of these matrices has been proposed as a method to select the optimal preprocessing methods that could lead to obtain better classification models during the data analysis stage, to distinguish sweet and bitter almonds and potentially other applications. In addition, the effect of instrumental and replacement replicates on the data sets was studied through the error covariance matrices. It was concluded that, across all devices, the instrumental variations had an insignificant impact on the measurements when compared to the variations arising from sample repositioning, rendering them negligible.

Furthermore, the proposed PLS-DA models successfully classified the almonds, this is, after applying the optimal preprocessing determined through the visual inspection of covariance and correlation matrices. All three spectrometers: SCiO, FlameNIR+ and NeoSpectra showed to be capable of classifying whole almonds by bitterness, with FlameNIR + performing better.

Funding

Grant URV Martí i Franqués – Banco Santander (2021PMF-BS-12). Grant PID2019-104269RR-C33 funded by MICIU/AEI/10.13039/501100011033. Chemometrics and Sensorics for Analytical Solutions (CHEMOSENS, ref.2021 SGR 00705, Departament de Recerca i Universitats, Generalitat de Catalunya). Financial support from the Spanish Ministry of Science, Innovation and Universities (MICIU) and the State Research Agency (AEI) (PID2019-106862RB-I00/AEI/10.13039/501100011033, PDC2021-120921-I00).

CRediT authorship contribution statement

Jokin Ezenarro: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jordi Riu:** Writing – review & editing, Validation, Resources, Methodology, Investigation, Conceptualization. **Hawbeer Jamal Ahmed:** Writing – original draft, Methodology, Investigation. **Olga Busto:** Funding acquisition. **Barbara Giussani:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Ricard Boqué:** Writing – review & editing, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

During the preparation of this work the authors used ChatGPT 3.5 in order to edit text and improve readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2024.126271.

References

[1] K.B. Beć, J. Grabska, C.W. Huck, Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives, Foods 11 (2022) 1465, https://doi.org/10.3390/FOODS11101465. Page 1465 11 (2022.
[2] H. Cen, Y. He, Theory and application of near infrared reflectance spectroscopy in determination of food quality, Trends Food Sci. Technol. 18 (2007) 72–83, https://doi.org/10.1016/J.TIFS.2006.09.003.
[3] L. Rodriguez-Saona, D.P. Aykas, K.R. Borba, A. Urtubia, Miniaturization of optical sensors and their potential for high-throughput screening of foods, Curr. Opin. Food Sci. 31 (2020) 136–150, https://doi.org/10.1016/J.COFS.2020.04.008.
[4] L.E. Agelet, C.R. Hurburgh, A tutorial on near infrared spectroscopy and its calibration, Crit. Rev. Anal. Chem. 40 (2010) 246–260, https://doi.org/10.1080/10408347.2010.515468.
[5] K.B. Beć, J. Grabska, C.W. Huck, Principles and applications of miniaturized near-infrared (NIR) spectrometers, Chem. Eur J. 27 (2021) 1514–1532, https://doi.org/10.1002/CHEM.202002838.
[6] Y. Ozaki, S. Šašić, J.H. Jiang, How can we unravel complicated near infrared spectra?—recent progress in spectral analysis methods for resolution enhancement and band assignments in the near infrared region, J. Near Infrared Spectrosc. 9 (2001) 63–95, https://doi.org/10.1255/JNIRS.295.
[7] Y. Ozaki, S. Morita, Y. Du, Spectral analysis, in: Near-Infrared Spectroscopy in Food Science and Technology, John Wiley & Sons, Ltd, 2006, pp. 47–72, https://doi.org/10.1002/9780470047705.CH3.
[8] J. Riu, G. Gorla, D. Chakif, R. Boqué, B. Giussani, Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis, Foods 9 (2020) 1090, https://doi.org/10.3390/FOODS9081090. Page 1090 9 (2020.
[9] B. Giussani, A.T. Escalante-Quiceno, R. Boqué, J. Riu, Measurement strategies for the classification of edible oils using low-cost miniaturised portable NIR instruments, Foods 10 (2021) 2856, https://doi.org/10.3390/FOODS10112856/S1.
[10] J. Riu, A. Vega, R. Boqué, B. Giussani, Exploring the analytical complexities in insect powder analysis using miniaturized NIR spectroscopy, Foods 11 (2022) 3524, https://doi.org/10.3390/FOODS11213524/S1.
[11] G. Gullifa, L. Barone, E. Papa, A. Giuffrida, S. Materazzi, R. Risoluti, Portable NIR spectroscopy: the route to green analytical chemistry, Front. Chem. 11 (2023), https://doi.org/10.3389/FCHEM.2023.1214825.
[12] B. Giussani, G. Gorla, J. Riu, Analytical chemistry strategies in the use of miniaturised NIR instruments: an overview, Crit. Rev. Anal. Chem. (2022), https://doi.org/10.1080/10408347.2022.2047607.
[13] J. Huang, Q. Wen, Q. Nie, F. Chang, Y. Zhou, Z. Wen, Miniaturized NIR spectrometer based on novel MOEMS scanning tilted grating, Micromachines 9 (2018) 478, https://doi.org/10.3390/MI9100478. Page 478 9 (2018).
[14] L.P. Schuler, J.S. Milne, J.M. Dell, L. Faraone, MEMS-based microspectrometer technologies for NIR and MIR wavelengths, J. Phys. D Appl. Phys. 42 (2009) 133001, https://doi.org/10.1088/0022-3727/42/13/133001.
[15] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36, https://doi.org/10.1016/J.ACA.2018.04.004.
[16] K.B. Beć, J. Grabska, H.W. Siesler, C.W. Huck, Handheld near-infrared spectrometers: where are we heading? NIR News 31 (2020) 28–35, https://doi.org/10.1177/0960336020916815.
[17] C.G. Kirchler, C.K. Pezzei, K.B. Beć, S. Mayr, M. Ishigaki, Y. Ozaki, C.W. Huck, Critical evaluation of spectral information of benchtop vs. portable near-infrared spectrometers: quantum chemistry and two-dimensional correlation spectroscopy for a better understanding of PLS regression models of the rosmarinic acid content in Rosmarini folium, Analyst 142 (2017) 455–464, https://doi.org/10.1039/C6AN02439D.
[18] V. Wiedemair, D. Mair, C. Held, C.W. Huck, Investigations into the use of handheld near-infrared spectrometer and novel semi-automated data analysis for the determination of protein content in different cultivars of Panicum miliaceum L, Talanta 205 (2019) 120115, https://doi.org/10.1016/J.TALANTA.2019.120115.
[19] U. Hoffmann, F. Pfeifer, C. Hsuing, H.W. Siesler, Spectra transfer between a fourier transform near-infrared laboratory and a miniaturized handheld near-infrared spectrometer, Appl. Spectrosc. 70 (2016) 852–860, https://doi.org/10.1177/0003702816638284.
[20] H. Yan, H.W. Siesler, Quantitative analysis of a pharmaceutical formulation: performance comparison of different handheld near-infrared spectrometers, J. Pharm. Biomed. Anal. 160 (2018) 179–186, https://doi.org/10.1016/J.JPBA.2018.07.048.
[21] B. Giussani, A.T. Escalante-Quiceno, R. Boqué, J. Riu, Measurement strategies for the classification of edible oils using low-cost miniaturised portable NIR instruments, Foods 10 (2021) 2856, https://doi.org/10.3390/FOODS10112856/S1.
[22] J. Riu, G. Gorla, D. Chakif, R. Boqué, B. Giussani, Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis, Foods 9 (2020) 1090, https://doi.org/10.3390/FOODS9081090. Page 1090 9 (2020.
[23] H. Yan, H.W. Siesler, Identification performance of different types of handheld near-infrared (NIR) spectrometers for the recycling of polymer commodities, Appl. Spectrosc. 72 (2018) 1362–1370, https://doi.org/10.1177/0003702818777260.
[24] J.A.F. Pierna, P. Vermeulen, B. Lecler, V. Baeten, P. Dardenne, Calibration transfer from dispersive instruments to handheld spectrometers, Appl. Spectrosc. 64 (2010) 644–648, https://doi.org/10.1366/000370210791414353.
[25] E. Zamora-Rojas, D. Pérez-Marín, E. De Pedro-Sanz, J.E. Guerrero-Ginel, A. Garrido-Varo, Handheld NIRS analysis for routine meat quality control: database transfer from at-line instruments, Chemometr. Intell. Lab. Syst. 114 (2012) 30–35, https://doi.org/10.1016/J.CHEMOLAB.2012.02.001.
[26] G. Gorla, P. Taborelli, C. Alamprese, S. Grassi, B. Giussani, On the importance of investigating data structure in miniaturized NIR spectroscopy measurements of food: the case study of sugar, Foods 12 (2023) 493, https://doi.org/10.3390/FOODS12030493/S1.
[27] G. Gorla, P. Taborelli, H.J. Ahmed, C. Alamprese, S. Grassi, R. Boqué, J. Riu, B. Giussani, Miniaturized NIR spectrometers in a nutshell: shining light over sources of variance, Chemosensors 11 (2023) 182, https://doi.org/10.3390/CHEMOSENSORS11030182/S1.
[28] G. Gorla, A. Taiana, R. Boqué, P. Bani, O. Gachiuta, B. Giussani, Unravelling error sources in miniaturized NIR spectroscopic measurements: the case study of forages, Anal. Chim. Acta 1211 (2022) 339900, https://doi.org/10.1016/J.ACA.2022.339900.
[29] P.D. Wentzell, C.C. Wicks, J.W.B. Braga, L.F. Soares, T.C.M. Pastore, V.T. R. Coradin, F. Davrieux, Implications of measurement error structure on the visualization of multivariate chemical data: hazards and alternatives, Can. J. Chem. 96 (2018) 738–748, https://doi.org/10.1139/cjc-2017-0730.
[30] G. Gorla, P. Taborelli, B. Giussani, A multivariate analysis-driven workflow to tackle uncertainties in miniaturized NIR data, Molecules 28 (2023) 7999, https://doi.org/10.3390/MOLECULES28247999. Page 7999 28 (2023.
[31] F. Matinrad, M. Kompany-Zareh, N. Omidikia, M. Dadashi, Systematic investigation of the measurement error structure in a smartphone-based spectrophotometer, Anal. Chim. Acta 1129 (2020) 98–107, https://doi.org/10.1016/J.ACA.2020.06.066.
[32] P.D. Wentzell, Measurement errors in multivariate chemical data, J. Braz. Chem. Soc. 25 (2014) 183–196, https://doi.org/10.5935/0103-5053.20130293.
[33] M.N. Leger, L. Vega-Montoto, P.D. Wentzell, Methods for systematic investigation of measurement error covariance matrices, Chemometr. Intell. Lab. Syst. 77 (2005) 181–205, https://doi.org/10.1016/J.CHEMOLAB.2004.09.017.
[34] M. Vega-Castellote, D. Pérez-Marín, I. Torres, J.M. Moreno-Rojas, M.T. Sánchez, 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, J. Food Eng. 294 (2021) 110406, https://doi.org/10.1016/J.JFOODENG.2020.110406.
[35] I. Torres, M.T. Sánchez, M. Vega-Castellote, D. Pérez-Marín, Fraud detection in batches of sweet almonds by portable near-infrared spectral devices, Foods 10 (2021), https://doi.org/10.3390/FOODS10061221, 1221 10 (2021) 1221.
[36] V. Wiedemair, D. Langore, R. Garsleitner, K. Dillinger, C. Huck, Investigations into the performance of a novel pocket-sized near-infrared spectrometer for cheese analysis, Molecules 24 (2019) 428, https://doi.org/10.3390/MOLECULES24030428. Page 428 24 (2019.
[37] C.W. Huck, New trend in instrumentation of NIR spectroscopy-miniaturization, in: Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications, Springer Singapore, 2020, pp. 193–210, https://doi.org/10.1007/978-981-15-8648-4_8.
[38] Ministerio de Agricultura Pesca y Alimentación, Frutos secos: Análisis de la realidad productiva 2021, 16–18, https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/frutas-y-hortalizas/Analisis%20realidad%20productiva%20frutos%20de%20cascara.aspx, 2022. (Accessed 2 October 2023).
[39] E. Borràs, J.M. Amigo, F. Van Den Berg, R. Boqué, O. Busto, Fast and robust discrimination of almonds (Prunus amygdalus) with respect to their bitterness by using near infrared and partial least squares-discriminant analysis, Food Chem. 153 (2013) 15–19, https://doi.org/10.1016/j.foodchem.2013.12.032.

[40] Å. Rinnan, F. Van Den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Anal. Chem. 28 (2009) 1201–1222, https://doi.org/10.1016/j.trac.2009.07.007.
[41] H. Büning-Pfaue, Analysis of water in food by near infrared spectroscopy, Food Chem. 82 (2003) 107–115, https://doi.org/10.1016/S0308-8146(02)00583-6.
[42] F. Allegrini, A.C. Olivieri, Recent advances in analytical figures of merit: heteroscedasticity strikes back, Anal. Methods 9 (2017) 739–743, https://doi.org/10.1039/c6ay02916g.