From spectroscopic data variability to optimal preprocessing: leveraging multivariate error in almond powder adulteration of different grain size

Barbara Giussani[1*], Manuel Monti[1], Jordi Riu[2]

[1]*Dipartimento di Scienza e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio 9, 22100 Como, Italy*

[2]*Universitat Rovira i Virgili. Department of Analytical Chemistry and Organic Chemistry. Carrer Marcel·lí Domingo 1, 43007 Tarragona, Spain*

* barbara.giussani@uninsubria.it

Abstract

Analysing samples in their original form is increasingly crucial in analytical chemistry due to the need for efficient and sustainable practices. Analytical chemists face the dual challenge of achieving accuracy while detecting minute analyte quantities in complex matrices, often requiring sample pretreatment. This necessitates the use of advanced techniques with low detection limits, but the emphasis on sensitivity can conflict with efforts to simplify procedures and reduce solvent use. This article discusses the shift toward green analytical methods, focusing on portable spectroscopic techniques in the near-infrared (NIR) region.

A case study involving the prediction of adulteration in almond flour with bitter almond flour illustrates the importance of particle size and the integration between the sample and the instrument. The study emphasizes the necessity of investigating the multivariate error associated with raw data to enhance data preprocessing strategies. This research provides valuable insights for professionals in the field, presenting a methodology applicable to a broad range of analytical applications while underscoring the critical role of raw data analysis in achieving accurate and reliable results.

Keywords

Multivariate measurements error; portable NIR sensors; almond; grain size; data preprocessing

1. Introduction

The challenge of analysing samples in their original form is becoming increasingly urgent in the field of analytical chemistry. As industries and regulatory bodies demand more efficient and sustainable practices, there is a growing emphasis on green analytical methods that minimize environmental impact [1, 2]. For instance, one crucial aspect of this transition involves reducing the use of toxic solvents, since these solvents can be harmful to both the environment and human health [3]. By moving toward greener analysis, chemists aim to develop methodologies that not only provide accurate results but also align with sustainability goals.

Moreover, the ability to use analytical techniques on-site is gaining importance. This approach significantly shortens the analysis chain, eliminating the need for sample transportation, storage, and the associated downtime. On-site analysis can lead to faster decision-making and more immediate responses to quality control issues, particularly in industries such as food safety, pharmaceuticals, and environmental monitoring [4–8]. However, this shift towards rapid, in situ analysis poses its own challenges.

While striving for efficiency, analytical chemists must also contend with the increasing demand for accuracy in their measurements. The need to detect very small quantities of analytes in complex matrices often necessitates pretreating samples to separate the analyte of interest from the matrix. This process requires the use of advanced analytical techniques capable of achieving exceptionally low detection limits to perform measurements. However, this focus on sensitivity can sometimes conflict with the goal of simplifying procedures and reducing reliance on solvents. The first principle of green analytical chemistry suggests eliminating sample preparation [9]. Notwithstanding, removing this step is almost impossible for complex samples and when high sensitivity is needed. Thus, instead of omitting or neglecting this step, efforts must be placed in adopting a framework for green sample preparation [10].

These two objectives—enhancing the efficiency of analytical processes and improving the accuracy of measurements—represent two sides of the same coin in analytical chemistry. Balancing these competing demands is one of the core challenges for modern analytical chemists, who must innovate and adapt to meet the evolving needs of their fields while adhering to the principles of sustainability and safety. As the discipline advances, the integration of green chemistry principles with cutting-edge analytical techniques will be essential in addressing these challenges and ensuring that analytical chemistry continues to play a vital role in various industries.

This article will focus on techniques that can be applied directly to the sample, specifically portable spectroscopic methods in the NIR region. The data processing for these techniques is conducted using chemometric methods. This combination aligns perfectly with the principles of rapid analysis that requires no solvents, making them compatible with green chemistry practices. However, it is important to note that recording the signal directly from the sample has its limitations, as the signal is influenced not only by the chemical composition of the sample but also by its physical characteristics. Therefore, data treatment also involves the crucial step of data preprocessing. Preprocessing is essential for optimizing the signal before modelling, making it more suitable for subsequent analysis [11].

The proper implementation of data preprocessing necessitates expertise and a thorough understanding of both the analytical data under investigation and the sample itself. While experienced chemometricians are typically well-acquainted with the most suitable preprocessing techniques for a given dataset, individuals who are new to chemometrics may find themselves uncertain about the best approaches to take. Often, preprocessing steps are applied without clear logic or rigorous reasoning. And frequently, the knowledge of the analytical data is either incomplete or entirely lacking, especially when it comes to new instruments or those that appear easy to use. This oversight can even occur among the most diligent scientists. The topic of preprocessing is prominently emphasized in the literature [12–16], and new methods for optimizing these techniques continue to emerge even today [17].

This article aims to focus specifically on the study of data generated by portable NIR spectroscopic instruments and how this analysis can primarily assist in identifying the sources of variability in both the sensors and in the experiment. Additionally, it can guide the selection of the best preprocessing techniques to achieve optimal chemometric models. The examination of raw data is sometimes minimized or entirely overlooked, but in our opinion, it is a crucial step in understanding the nature of the data and determining how to preprocess it for more effective modelling.

To achieve this, a case study of relevance in the food sector, though its framework can be extended to many other samples and sectors, was chosen. It focuses on the possibility of predicting the adulteration of almond flour of different particle sizes with bitter almond flour (which is not suitable for human consumption [18]), using portable NIR sensors. These miniaturized sensors are making inroads into the portable instrumentation market, with all their pros and cons [19]. In a previous study, it has been demonstrated that portable NIR sensors are capable of distinguishing whole sweet almonds from bitter ones [20]. Building on this result, an effort was made to take it a step further by quantifying possible contamination in a granular sample while studying the potential influence of grain size. Since different NIR sensors may present very different technical and instrumental configurations [21], the grain size of the target sample may be an important factor to influence the performance of the measurements obtained with a particular NIR instrument. The focus will be on analysing the raw signal through an assessment of the multivariate error associated with the data. Examining the behaviour of this error can provide insights into how to improve the experiment and how to appropriately preprocess the raw data to make it more suitable for Partial Least Squares regression [22–25], the most widely multivariate regression method used in the literature.

2. Materials and methods

2.1 NIR miniaturized sensors

Five measurement strategies were employed to analyse almond powder using four different miniaturized NIR instruments:

1. SCiO (Consumer Physics, Herzliya, Israel): this compact NIR device measures 67.7 mm x 40.2 mm x 18.8 mm and weighs 35 g. It operates in the wavelength range of 740 to 1070 nm, controlled via an Android smartphone using the 'SCiO Lab' app over Bluetooth. With a default scan time of 2 to 5 seconds, spectra are stored in the cloud and can be downloaded from ‘The Lab’ website. Calibration is necessary before initial measurements using a reference standard located on the back cover of the instrument.

2. NeoSpectra Microdevelopment Kit (MDK) (Si-Ware, Cairo, Egypt): two NeoSpectra MDK units, referred to in this text as MDK1 and MDK2, were used. The instruments measure 32 mm x 32 mm x 22 mm and with a weight of 17 g. They utilize a micro-electromechanical system (MEMS) Michelson interferometer and operate between 1350 to 2558 nm with a resolution of 16 nm. They connect to a Raspberry Pi, allowing configuration of scan time and run modes. Calibration is required each time the software is initiated, and the scan time was optimized to 5 seconds without data interpolation.

3. NeoSpectra Scanner (Si-Ware, Cairo, Egypt): this device measures 180 mm x 45 mm x 8 mm and weighs 730 g, operating within 1351 to 2559 nm. It also features a MEMS Michelson interferometer and uses a proprietary mobile app for operation. With a scan time of 5 seconds without data interpolation, it can function on battery power. A Rotator accessory was also utilized to improve sample representativeness during measurements. The Rotator accessory attaches to one side of the NeoSpectra Scanner and enables automatic rotation of the sample positioned on top of the device, enhancing the representativeness of the area exposed for analysis. This accessory is particularly suitable for applications requiring the analysis of non-homogeneous samples without sample preparation.

Therefore, the five measurements strategies are SCiO, MDK1, MDK2, NeoSpectra Scanner and NeoSpectra Scanner with the rotator accessory sensors.

All instruments employed reflectance mode analysis, beginning with a background measurement using a 99% reflectance Spectralon® standard. The SCiO and NeoSpectra Scanner integrated this material into their covers. Analytical sessions for the NeoSpectra MDK sensors commenced after a 20-minute warm-up.

Spectra collection was conducted in contact mode between the instruments and the sample, or using a custom sample holder made with a 3D printer [26] which included a borosilicate glass coverslip at the intermediate bottom to help contain the sample. Fifteen replicates with sample repositioning were conducted to study the multivariate error. For the NeoSpectra MDK sensors, only 5 replicates were performed due to the longer analysis times: the use of the NeoSpectra MDK sensors involved the custom sample holder (containing the almond flour) described above in direct contact with the instrument's measurement window. For each analytical replicate, the sample holder was removed and reinserted into the instrument, and the optical window of the NeoSpectra MDK was cleaned before the next measurement. This additional handling slowed the measurement process. In contrast, measurements conducted using the SCiO or the NeoSpectra Scanner (with or without the Rotator accessory) were significantly more straightforward.

Only replicates with repositioning were carried out, not purely instrumental replicates, as previous studies have demonstrated that the latter have a much smaller impact on the quality of the collected data [20, 27].

2.2 Reference sample preparation

Whole sweet almonds (unskinned) were purchased from local markets in Tarragona, Spain, while whole bitter almonds (skinned) were obtained from Schmuetz Naturkost (Malente, Germany). To remove the skins from the bitter almonds, the almonds were briefly boiled in water for one minute, after which the skins were manually peeled off. The almonds were then air-dried for one day, maintaining their hardness and physical integrity.

Almond powder was produced using three different particle sizes (referred to as large, medium, and small in this article). These sizes were achieved by grinding the almonds in a coffee grinder (Black+Decker BXCG150e, Oliana, Spain), and then sieving the ground material through three mesh sizes (5 holes/inch, 12 holes/inch, and 22 holes/inch, Lacor 68342, Bergara, Spain). The grinding process was carefully optimized to prevent significant temperature increases, ensuring the chemical stability of the samples throughout the procedure. All almonds were ground in a single session to ensure uniform treatment. Each particle size was analysed using the five measurement strategies described in the previous section.

For each particle size, 20 reference samples were prepared with varying concentrations of bitter almond adulterant: 0%, 1%, 2%, 4%, 6%, 8%, 10%, 12%, 14%, 16%, 18%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 75%, and 100% by weight. The reference samples were labelled and stored in zip-lock bags in the refrigerator at 5°C, and they were allowed to reach room temperature before analysis.

2.3 Data analysis

2.3.1 Data arrangement

The spectroscopic data were organized into several matrices, with the rows representing the recorded spectra and the columns corresponding to the variables measured by the different sensors. Specifically, the various sensors include:

SCiO: 330 wavelengths

NeoSpectra MDK1: 142 wavelengths

NeoSpectra MDK2: 142 wavelengths

NeoSpectra Scanner: 257 wavelengths

It is worth noting that, although the two MDK sensors are of the same type and record the same number of wavelengths within the same range, the wavelengths are not identical. This implies that the two instruments are not truly directly comparable. This is one of the challenges with this type of portable instrumentation, which is still evolving, relatively new to the market, and undergoing optimization.

All replicates for each sample were incorporated into the matrices for the calculation of the error covariance and correlation. The average spectra were utilized for the regression model calculations.

2.3.2 Error covariance and correlation calculation

A prevalent approach for assessing multivariate measurement errors is using the error covariance matrix (ECM) [28]. The literature identifies three main methods for estimating this matrix: experimental replication, theoretical prediction, and empirical modelling. The experimental method allows for straightforward estimation by conducting calculations after acquiring an adequate number of sample replicates. The theoretical method, however, necessitates an in-depth understanding of potential error sources beforehand. Empirical modelling is between the two, leveraging both replicate measurements and theoretical insights.

When multiple replicate measurements are feasible, as in the case of measurements using portable spectroscopic sensors, the process of characterizing multivariate measurement errors begins by determining the "true" sample spectrum based on the number of replicates, denoted as *r*. The residuals matrix containing all the individual residuals $\mathbf{e}_i$, is calculated by subtracting the average spectrum ($\bar{\boldsymbol{x}}$) from each spectrum ($\boldsymbol{x}_i$). Subsequently, the error covariance matrix ($\boldsymbol{\Sigma}$) is determined through the covariance of the residuals, as illustrated in:

$$\boldsymbol{\Sigma} = \frac{\sum_{i=1}^{r} \mathbf{e}_i^{\mathrm{T}} \mathbf{e}_i}{(r-1)}$$

The error correlation matrix, which contains the correlation coefficients for the elements of $\boldsymbol{\Sigma}$, can be computed as follows

$$\boldsymbol{\Sigma}_{corr} = \boldsymbol{\Sigma} / \sqrt{diag(\boldsymbol{\Sigma}) \cdot diag(\boldsymbol{\Sigma})^T}$$

where *diag*($\boldsymbol{\Sigma}$) is the diagonal of the error covariance matrix. Variance estimates derived from experimental data often exhibit substantial uncertainty, largely due to experimental constraints that limit the number of replicates. To ensure more reliable estimations, it is important to either increase the number of replicates or consolidate error covariance across different subsets of samples. This averaging approach can be effective, particularly when the measurement data show minimal differences among samples from the same origin, such as in near-infrared spectra [29].

After obtaining the error covariance and correlation matrices, visual interpretation can be performed. The error covariance matrix reveals the relationships between measurement errors at various wavelengths. The diagonal elements indicate the uniformity of errors across the spectra (homoscedasticity), while off-diagonal values suggest inconsistency in errors (heteroscedasticity) [30]. The off-diagonal elements provide insight into the covariance of measurement errors, demonstrating the strength of these relationships. The error correlation matrix, which is derived from the covariance matrix, offers a structure of these relationships independent of scale, featuring values that range between -1 and 1.

An example of the representation of the error covariance matrix, its diagonal (the visualization of the diagonal of the covariance matrix simplifies in some cases the interpretation of the results), and the error correlation matrix is shown in Figure 1. Data in Figure 1 correspond to the analysis with the NeoSpectra Scanner sensor of the 15 replicates of the reference sample with medium grain size and 16% of bitter almond adulterant.

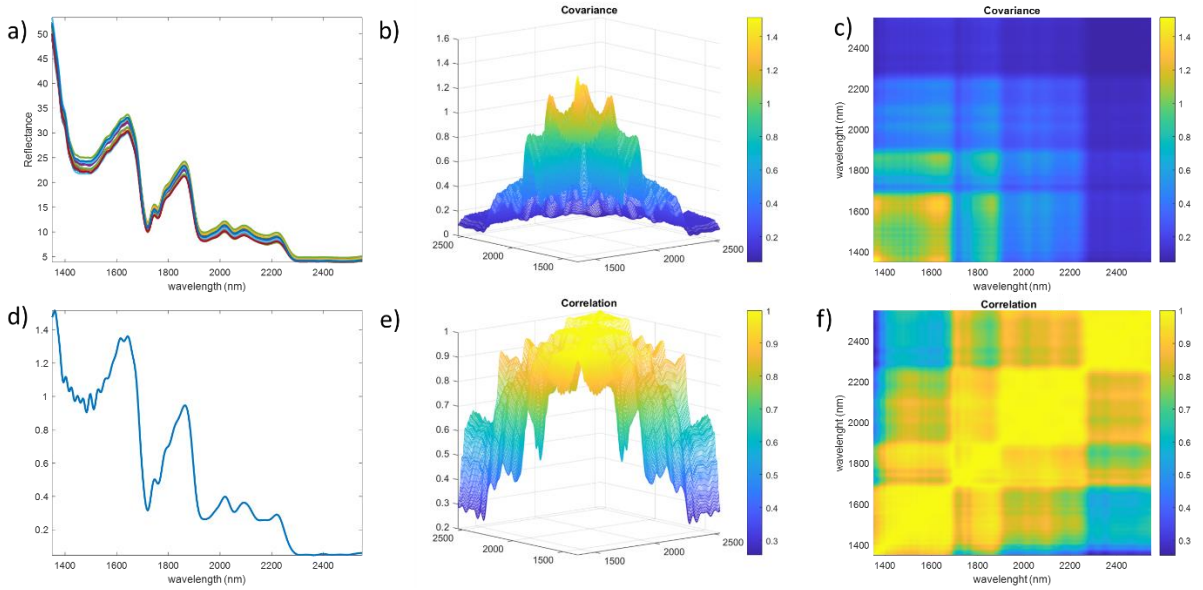


Fig. 1. a) Spectra of 15 replicates of medium grain size almond with 16% of bitter almond adulterant; b) error covariance matrix; c) error covariance matrix viewed from above; d) diagonal of the error covariance matrix; e) error correlation matrix; f) error covariance matrix viewed from above.

2.3.3 Partial least squares regression and data preprocessing

Fifteen distinct cases were generated from this study, comprising the five experimental configurations and the three particle sizes studied. Models were built with 10 reference samples constructed between 0% and 100% adulteration using bitter almond flour. 10 reference samples measured in an independent analytical session were used for external validation. The samples were randomly divided before the construction of the models. Extreme samples (those with the highest and lowest percentages of adulteration) were included in the model construction but excluded from the external validation. This was done to ensure that all validation samples fell within the range of percentages covered by the models.

Partial least squares (PLS) regression is a multivariate statistical method used to model relationships between multiple independent variables and one or more dependent variables. This method is particularly advantageous when predictors exhibit high collinearity as in the case of spectroscopic signals. PLS regression works by extracting latent variables from the original variables (wavelengths of the spectra in this case) that capture the maximum variance while also maximizing the covariance with the response variables (the % of bitter almond powder in this case). This approach allows for robust predictions and insights into the underlying data structure. Calculations were performed on averaged spectra, and, as previously mentioned, the models were validated using external validation [31]. The performance of the regression models obtained were evaluated using the prediction error computed with the test set values, often referred to as RMSEP (Root Mean Square Error of Prediction):

$$RMSEP = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

where y_i is the i^{th} sample of the $\boldsymbol{y}$ vector, $\hat{y}_i$ is the predicted value of the i^{th} sample using the PLS regression model, and n is the number of samples (each sample is the average spectra of the replicate measurements).

A lower RMSEP indicates better model performance, as it reflects that the values predicted by the model are closer to the reference values. Another important parameter to consider is the coefficient of determination (R^2), which measures the agreement between the model predicted outcomes and the reference results (the values from the ***y*** vector). The optimal model was chosen as the best balance between RMSEP (the lower possible) and the number of latent variables used. Models with fewer latent variables are more parsimonious and generally easier to interpret. The scree plot (RMSEP values versus the number of latent variables) was used in any case to determine the number of latent variables. The best model in each case was chosen as the best compromise between a lower root mean square error of prediction (RMSEP) and a smaller number of latent variables to avoid overfitting.

Very often, spectroscopic data, and even more so data collected via external reflection (as in the case study of this article), are not suitable for developing multivariate regression models in their raw form. They may contain, and almost always do contain, noise or scattering information that is not always beneficial for modelling. Therefore, various preprocessing methods can be utilized to enhance the PLS models. Standard spectra preprocessing methods were examined [11, 13], including Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), first and second Savitzky-Golay derivatives, as well as detrending, smoothing (moving window with different smoothing points) and baseline correction techniques.

Data elaboration and analysis were performed with in-house routines programmed in MATLAB R2024a (Mathworks Inc., Natick, MA, USA) and PLS toolbox 9.3.1 (Eigenvector Inc., Manson, WA, USA). All the models were finally mean centered.

3. Results and discussion

3.1 Errors affecting reference samples

This study aims to achieve the best prediction models possible through multivariate regression while preprocessing the data based on the results obtained from the multivariate error analysis.

In the construction of the PLS regression models, for each of the 20 reference samples of adulterated almond flour, we performed replicates with sample repositioning (15 replicates for the NeoSpectra scanner and the SCiO sensors, and 5 replicates for the NeoSpectra MDK sensors). The 20 average spectra recorded by the SCiO, NeoSpectra MDK1, and NeoSpectra Scanner sensors for the reference samples with small grain size are shown as an example in Figure 2. The average spectra for NeoSpectra MDK2 are visually similar to those for NeoSpectra MDK1, and the average spectra for the NeoSpectra Scanner Rotor are visually similar to those for NeoSpectra Scanner.

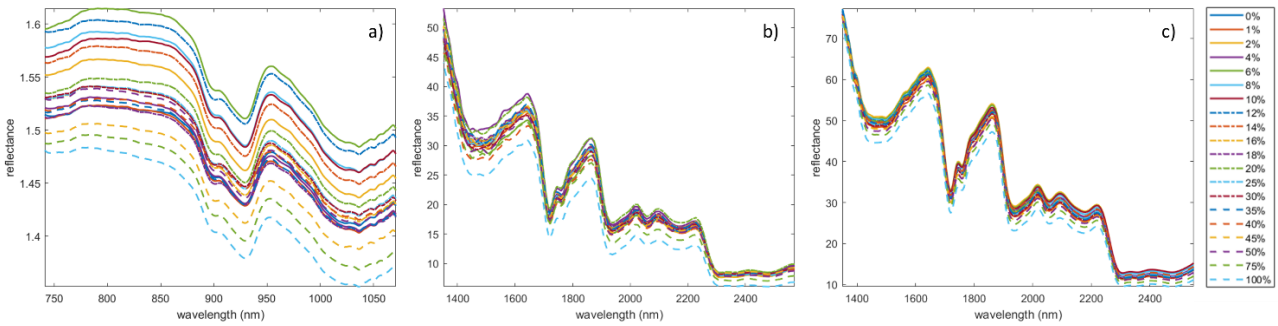


Fig. 2 Average spectra of all the reference samples - a) SCiO; b) NeoSpectra MDK1; c) NeoSpectra Scanner.

Replicates allowed us to investigate whether the dispersion of results varied for each reference sample or was comparable across the entire regression range. Variable dispersion of replicates across reference samples

indicates heteroscedasticity, which leads to inefficient regression estimates and unreliable statistical tests. This variation can disproportionately affect the model, complicating predictions and necessitating appropriate strategies, such as robust techniques or data transformations to stabilize the variance. Conducting diagnostic analyses is crucial for identifying and addressing these issues to enhance the reliability of the regression model.

The study of multivariate error in the raw data provides valuable information about the spectra. It is worth noting that, in a previous study, we demonstrated that the error associated with measurements of bitter and sweet almonds (the study was conducted on whole almonds) did not show any significant differences between the two types of almonds [20]. In the present work, it will be interesting to investigate whether varying percentages of bitter and sweet almonds in samples with different grain sizes affect the error in the spectra of the reference samples.

Data recorded for all the reference samples were used. It is worth noting that the study of multivariate error relates to the spectroscopic measurement and depends on the characteristics of the instrument, on the sample, and on the environment: in short, the experiment as a whole. Therefore, all samples, whether used for building the calibration model or for its validation, are valuable for this purpose. To simplify the analysis of the results in this initial phase of the study, the plots of the error covariance or correlation matrices were not examined: at this stage, it is sufficient to analyse the diagonal of the error covariance matrix. Thus, for each of the reference samples with a known percentage of adulteration, the multivariate error covariance matrix was calculated, and its diagonal was extracted (Figure 3). Previous studies have demonstrated that this is an effective and intuitive method for the initial interpretation of data [32, 33]. It is worth mentioning that replicate spectra significantly deviating from the typical signal were removed. In other words, spectra identified as outliers were excluded. These represent a very small portion of the total collected signals (0.11% of the spectra for the SCiO, 0.33% of the spectra for the NeoSpectra Scanner, and 1% of the spectra for the NeoSpectra Scanner with the Rotator accessory) and can be attributed to gross measurement errors.

Two aspects must be considered in this case: first, the number of replicates is different, making it challenging to compare the results obtained from different instruments. A balanced replication is also required when comparing the instruments. However, it is possible to compare the results obtained from the same instrument across different particle sizes. The second point is that the number of replicates is limited, which does not provide an excellent estimate of the error in the raw data, but it does give an indication of the presence of homoscedasticity in the data. The following section will discuss how to obtain a better estimate of the error on the raw data.

The diagonal of the covariance matrix for the 20 reference samples, which represents the 20 different percentages of adulteration, is shown in Figure 3. In the graphs in Figure 3, the *x*-axis represents the wavelengths measured by the instrument, and the *y*-axis shows the value of the diagonal of the error covariance matrix calculated for each reference sample. As can be observed across all the graphs, there is no discernible trend with respect to the percentage of adulteration: the curves are intermingled and do not display any pattern attributable to the level of adulteration. It can be observed in all the graphs that the variance value (recalling that the diagonal of the error covariance matrix represents the variance of the variables) is always greater than zero and tends to increase as the wavelength decreases. This trend provides interesting information that will be discussed in more detail in the next section.

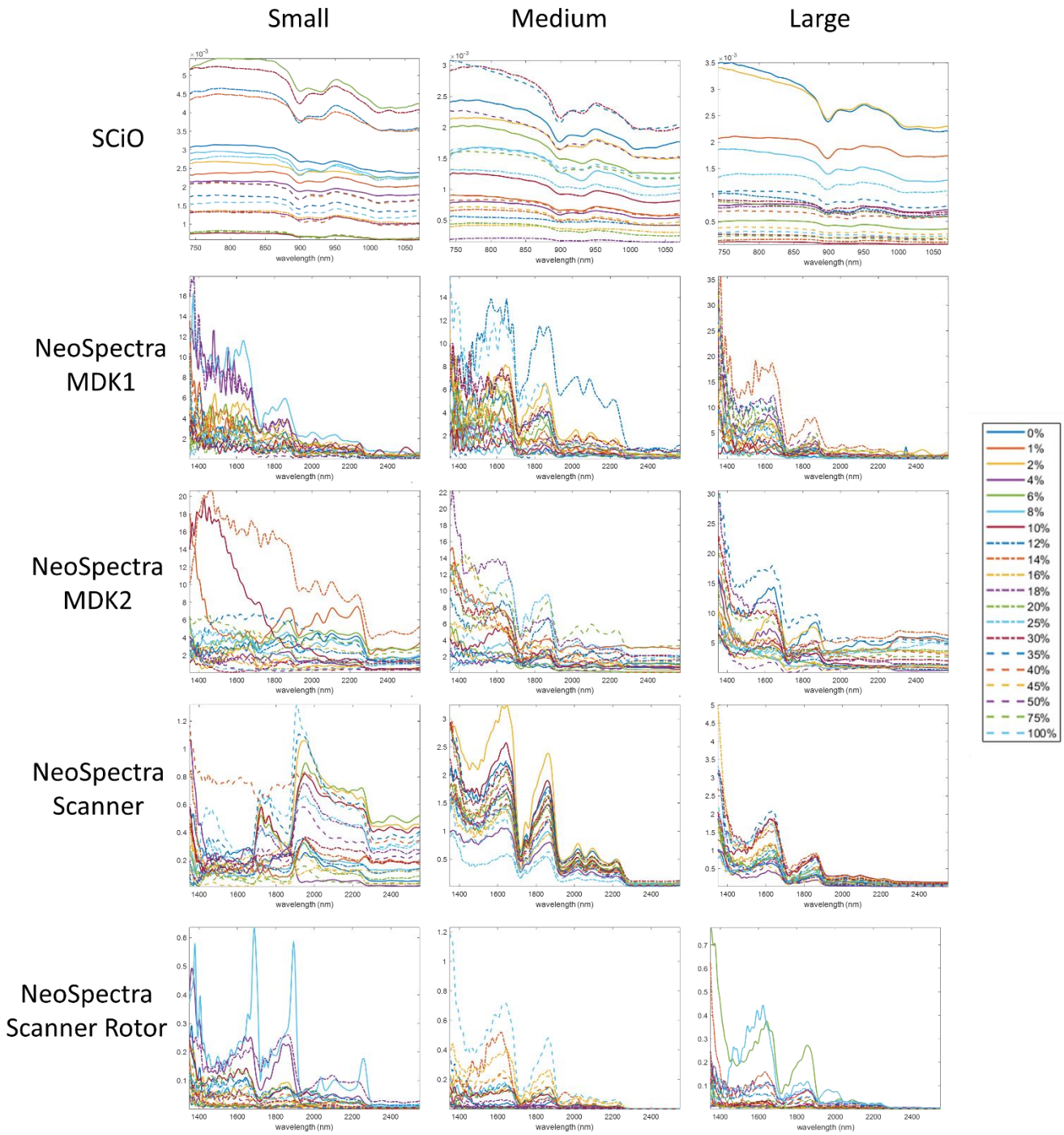


Fig. 3 Diagonal of the covariance matrix for the 20 reference samples for the three grain sizes and the five measurement strategies.

It is now worth focusing on the analysis of the results from the individual instruments. For the NeoSpectra MDK1 sensor, it is observed that the variance increases as the particle size increases: the graphs show larger values for larger particle sizes. The distinct particle size of the reference samples significantly affects the magnitude of the multivariate error in the raw data. A similar behaviour can also be observed for the NeoSpectra MDK2 sensor, though it is not identical. These two sensors, despite being of the same type and brand, do not exhibit identical behaviour. This underscores the importance of studying the raw data error of the specific spectrophotometer in detail for any given application. It is worth noting that only five replicates were available for these sensors, making it difficult at times to identify peculiar spectra. This may also cause significant variability in the error covariance matrices and their diagonals (e.g., MDK2 graph, smallest particle

size). Now, analysing the results obtained from the NeoSpectra Scanner sensor, it can be observed that the variance is significantly lower than that of the more affordable sensors (i.e. NeoSpectra MDK) from the same manufacturer. The smallest variance values are found for the smallest particle size, while the values for medium and large particle sizes are essentially the same. When the NeoSpectra Scanner sensor is used with the Rotator accessory, the error value decreases further, and there is no evident difference in the variance values across the three particle sizes. For the SCiO sensor, it is not possible to make a comparison with the other sensors from this perspective: it operates at different wavelengths with different raw reflection values, which prevents a numerical comparison of the variance values. However, it is possible to observe that the lowest variance values correspond to the largest particle size, and they increase as the particle size of the reference samples decreases.

These results are consistent with previously obtained results from sugar samples. In that case, granular sugar was compared to the same sample in solid form, as sugar lumps. In that instance as well, the physical characteristics of the investigated sample influenced the error associated with the measurements [33]. For almond flours with different grain sizes, the results vary for each sensor, and the findings from one sensor cannot be fully applied to another, even if both are from the same manufacturer. The variation in the data is likely also related to the spot size (approximately 1 mm for the NeoSpectra MDK, around 5 mm for the SCiO, and about 10 mm for the NeoSpectra Scanner); however, quantifying this influence is complicated.

In addition to being a system that assesses whether the standard samples we have selected are truly suitable for constructing a multivariate regression model from the perspective of homoscedasticity, which is a fundamental parameter, the approach proposed in this section is also an excellent way to globally evaluate the error associated with the raw data of a portable spectrophotometric sensor for a specific application. This is one of the aspects to consider when choosing the most suitable sensor for the application of interest.

3.2 Estimation and interpretation of the error in the raw data

After demonstrating through the analysis of the diagonal of the error covariance matrix that the error values related to the raw data are independent of the percentage of adulteration, these values can be used to better estimate the measurement error associated with this type of sample for each of the available miniaturized sensors. Consequently, information from the different reference samples at various percentages of adulteration can be aggregated (once again, both the calibration samples for the regression models and the validation samples can be used at this stage). However, information across the different particle sizes cannot be aggregated, as they have been shown to significantly influence the measurement error.

Figure 4 illustrates one of the ways in which error covariance and correlation matrices can be represented, specifically through a 3D graph. The covariance and correlation matrices of the error were calculated for each particle size and each sensor. A view from above of the error covariance and correlation matrices shown in Figure 4 can be found in Figure S2 of the Electronic Supplementary Material.

The focus will not be on comparing the numerical values between the different experimental configurations, as this has already been addressed in the previous section through the analysis of the diagonals of the error covariance matrices. This section will concentrate on the shape of the two matrices and the information that can be derived from them.

The primary effects that can be graphically identified in all the error covariance matrices include offset noise, shot noise, and multiplicative noise [28, 34]. The presence of offset noise is evident from the observation that the entire covariance structure consistently exhibits values greater than zero, indicating a systematic shift in the signal baseline. Moreover, the offset observed is proportional to the square root of the signal amplitude in a multiplicative manner, which categorizes it as a heteroscedastic error, commonly referred to as shot noise.

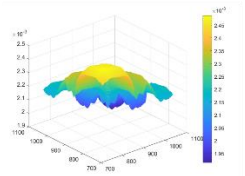
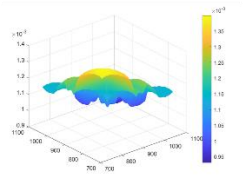
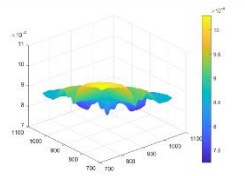
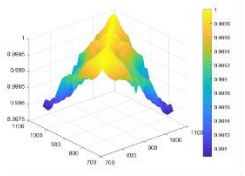
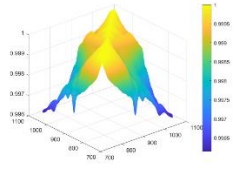
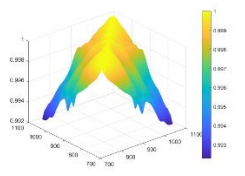
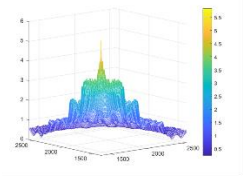
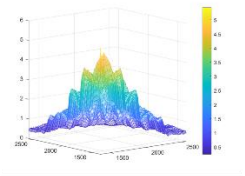
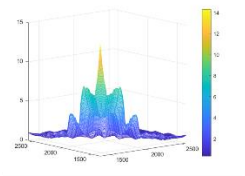
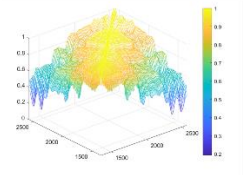
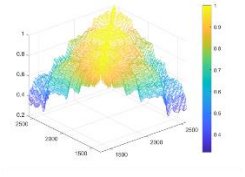
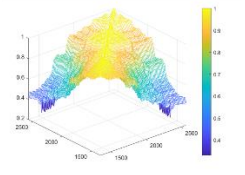
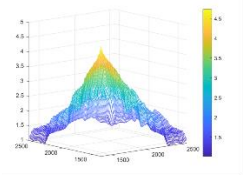
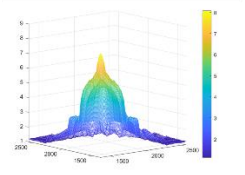
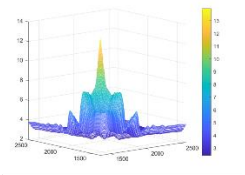
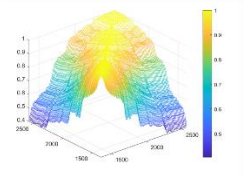
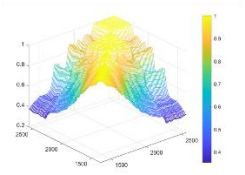
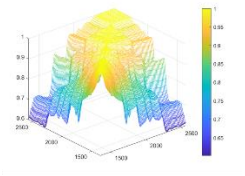
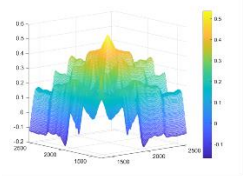
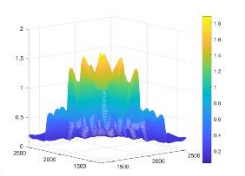
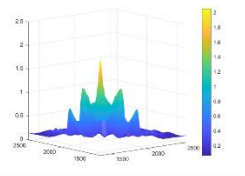
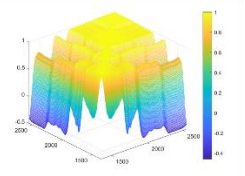
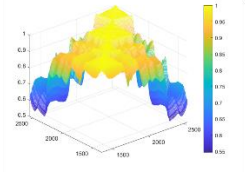
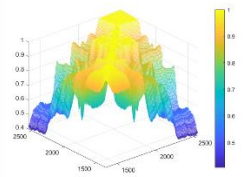
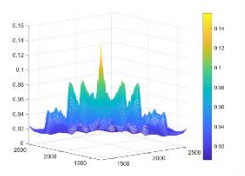
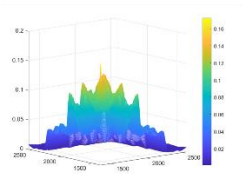
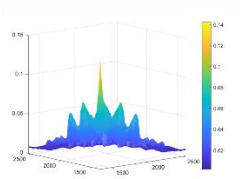
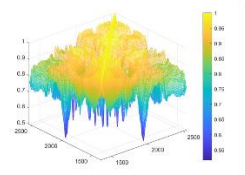
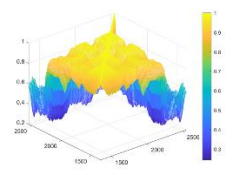
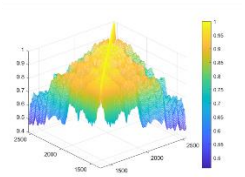
	Small	Medium	Large
SCiO error covariance matrix			
SCiO error correlation matrix			
NeoSpectra MDK1 error covariance matrix			
NeoSpectra MDK1 error correlation matrix			
NeoSpectra MDK2 error covariance matrix			
NeoSpectra MDK2 error correlation matrix			
NeoSpectra Scanner error covariance matrix			
NeoSpectra Scanner error correlation matrix			
NeoSpectra Scanner Rotor error covariance matrix			
NeoSpectra Scanner Rotor error correlation matrix			

Fig. 4 Error covariance matrices and error correlation matrices for the small, medium and large particle sizes (in columns) and the five measurement strategies tested (in rows).

In addition to these, multiplicative noise can be discerned through the increasing covariance error structure, which displays a relationship with the mean reflectance of the signals. As the mean reflectance diminishes, the variability in the error structure escalates, indicating that the noise characteristics become more pronounced at lower reflectance levels. This interaction underscores the complexity of the noise phenomena present in the data and highlights the necessity for careful interpretation and analysis in spectroscopic measurements. Understanding these noise types is crucial for refining measurement techniques and improving the overall accuracy and reliability of the results.

For the NeoSpectra MDK1 and MDK2 sensors, the correlation matrices show a very high correlation in different regions: below 1700 nm, around 1800 nm, and above 1900 nm. The degree of correlation outside these regions is lower. This behaviour can be explained by the presence of multiplicative and constant errors. For the NeoSpectra Scanner sensor, the correlation matrices reveal a strong correlation in four distinct regions for medium and large particle sizes: below 1700 nm, around 1800 nm, between 1900 and 2300 nm, and above 2300 nm. The correlation outside these regions is noticeably lower, likely due to the presence of multiplicative and constant errors. In contrast, for small particle sizes, a high correlation is observed across nearly the entire wavelength range. When the sensor is paired with the Rotator accessory, the correlation matrices show a very high correlation in different regions. For medium and large particle sizes, the correlation is very strong for wavelengths below 2300 nm. For small particle sizes, a high correlation is observed in several regions: around 1500 nm, around 1800 nm, between 2000 nm and 2200 nm, and above 2250 nm. This pattern can be attributed to the presence of both multiplicative and constant errors. In the case of the SCiO sensor, due to the different scale, it is challenging to compare the results with the other instruments, as previously commented. The error structure for the SCiO is quite similar across the three particle sizes. In all three cases, a constant error (offset) and a multiplicative error, directly proportional to the reflectance values, can be observed. A high degree of correlation between all variables is evident, as indicated by the magnitudes in the correlation matrix.

The analysis of the error matrices provides insight into the key characteristics of these data, helping to guide the selection of the most appropriate preprocessing methods. In this case, they will likely need to correct for baseline shifts and multiplicative effects, likely due to scattering. It is important to note, once again, that while all the instruments exhibit these characteristics in their data, the intensity and characteristics of these effects vary. This variation can be an important consideration when selecting the best sensor for a specific application.

3.3 Using the information provided by multivariate error to develop the best regression model

Analysing samples without any sample pretreatment, with inherently multivariate techniques such as spectroscopic methods, offers numerous advantages. However, it often requires an appropriate level of expertise in data analysis, as the data typically need to be preprocessed before being used in multivariate regression models. This highlights the dual nature of analytical chemistry: one can pretreat samples to extract the analyte of interest, thereby simplifying the analysis and subsequent data preprocessing. Alternatively, one can analyse the samples in their raw form, which means that the more complex aspect of the analytical methodology will be the study and interpretation of the resulting data to extract the relevant information.

At present, many guides are available for properly preprocessing chemical data (some of which are among the references of this article). The approach proposed in this article is not intended as an alternative to those

methods and guidelines, but rather as a supplementary perspective to address the issue. This method offers a complementary strategy based on a deep understanding of the errors associated with the data of interest, guiding scientists in selecting the most appropriate approach. The strength of this method lies in its visual nature, allowing for valuable insights to be drawn from graphs instead of equations or tables of numbers. This is a well-known advantage of chemometric techniques, which offer users, as well as decision-makers relying on their results, an easily interpretable visual framework based on two- or three-dimensional graphs.

As an example, the study of the best PLS models for each of the portable NIR sensors used in this work is presented in Figure 5. The results for all 15 calculated PLS models are summarized in Table S1 in the Electronic Supplementary Material. Furthermore, Figure S1 in the Electronic Supplementary Material presents an example of the predicted versus measured values for the validation set of the PLS model obtained using the NeoSpectra with the Rotator accessory for small grain size measurements, which represents the best PLS model among the 15 calculated. In this case, we study the error correlation matrices and visualize them using 2D visualization, which makes it easier to capture their characteristics. Starting from the raw data and attempting to correct the errors suggested by the multivariate error study, notably good models were obtained, with prediction errors comparable to those found in the literature.

Figure 5 shows how the error correlation matrix changes after applying the best sequence of data preprocessing. From these figures, important findings emerge. The first one is that the correlation between errors (off-diagonal values) decreases after applying preprocessing. The second is that the preprocessing methods used have indeed improved the data, making it more suitable for modelling. Before applying the preprocessing, the errors exhibited a structured pattern, while after preprocessing, the distribution of errors appeared less structured and more random, thus better aligned with the requirements of Partial Least Squares (PLS) modelling, which ideally seeks a random error distribution in the off-diagonal values. This insight can be gleaned from the observation of the error correlation matrix.

The best model for predicting the adulteration of sweet almond flour with bitter almond was obtained using the NeoSpectra Scanner equipped with the Rotator accessory, on the small particle size sample, achieving a prediction error (RMSEP) of 3.8% with 3 latent variables and using data preprocessing that corrected multiplicative scattering (MSC) and applied first derivative, that emphasizes changes in the data, making differences between samples more apparent and enhancing the separation of spectral peaks. The two NeoSpectra MDK1 and MDK2 sensors performed best on small particle size samples, with RMSEP values of 9.3% and 6.4%, respectively, again with 3 latent variables. In this case, a smoothing of the data was implemented before MSC and the second derivative, as the data appeared to be noisier than those from the other instruments. It is worth noting that, as mentioned earlier, although these two instruments are technically of the same type, they deliver different performances, corresponding to variations in the multivariate error associated with the measurement. As expected, the SCiO provided the best results for the large particle size, with an RMSEP of 4.5% (4 latent variables). In the case of SCiO, no smoothing was needed, and the data preprocessing used corrected the multiplicative scattering (MSC) and emphasized changes in data (first derivative). Despite the preprocessing effectively enhances the original data and leads to good prediction models, the overall error of the model continues to follow the trend observed when analysing the raw data. This means that the best models are derived from the combination of instrument and particle size that yielded the lowest error on the raw spectroscopic data. Once again, the analysis of raw data comparing different sensors can aid in the selection of a sensor for a specific application.

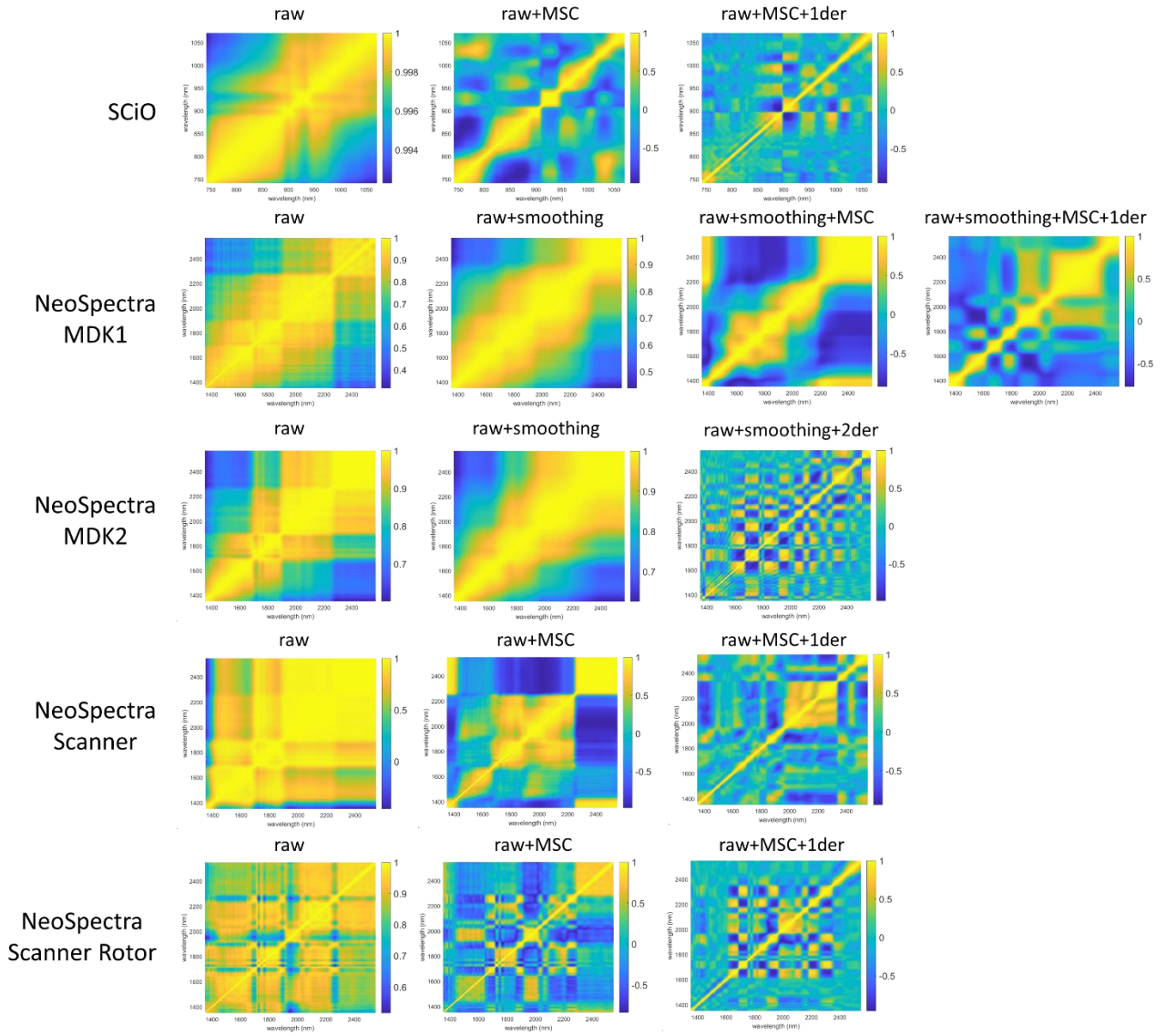


Fig. 5 Error correlation matrix for the five measurement strategies with the best combination of preprocessing methods applied. SCiO: large particle size, all other sensors: small particle size. MSC: multiplicative scatter correction. Smoothing: moving window smoothing (17 points for all the grain sizes in NeoSpectra MDK1, 13 points for small and medium sizes in NeoSpectra MDK2, 11 points for large size in NeoSpectra MDK2). 1der: Savitzky-Golay first order derivative. 2der: Savitzky-Golay second order derivative.

We can also make another observation regarding the models obtained. The NeoSpectra Scanner, whether equipped with the Rotator accessory or not, demonstrates relatively consistent prediction errors for adulteration across the three particle sizes in the validation set, ranging from 3.8% to 5.6%. In contrast, the situation worsens for the NeoSpectra MDK sensors, where the RMSEP for the MDK1 sensor varies between 9.3% and 13%, and between 6.4% and 15% for the MDK2 sensor (the highest values were obtained for the large particle size). This indicates that changing the particle size of the sample leads to significant variations in the RMSEP values. This finding aligns with what emerged from the study of multivariate error in the raw data. The SCiO sensor showed the highest variability, with its least effective model for small particle size yielding an RMSEP of 13%. All these results are shown in Table S1 of the Electronic Supplementary Material.

4. Conclusions

This study underscores the importance of investigating the error associated with raw data, particularly in the context of spectroscopic techniques, especially those that are relatively new among the various analytical methods available. It presents a methodology that can be effectively utilized by researchers in this field and beyond.

The innovative approach showcased here focuses on determining the adulteration of almond flours with varying particle sizes, a method that can be easily extended to any analytical issue involving powder samples. It highlights the significance of particle size in these studies and emphasizes the necessity of examining the compatibility and synergy between sample characteristics and instrument capabilities to achieve accurate and reliable analytical results. Consequently, we propose studying the raw data through the examination and interpretation of multivariate error.

The prediction errors obtained from the PLS modelling are very promising, considering the explored range. In future work, even more improved models could likely be achieved by concentrating calibration efforts on the lower end of the adulteration values; however, this was not the primary aim of our study. To determine whether the adulteration of bitter almonds influenced the multivariate error and to assess homoscedasticity in the calibration, we included values that encompassed the entire range of variability.

While preprocessing enhances the data, making it more suitable for multivariate modelling, the results ultimately depend on the quality of the initial data. Therefore, it is crucial to conduct thorough analyses. Investigating multivariate error provides insights that can guide the data preprocessing procedures and help identify the most effective sensor for a given sample. This finding is particularly beneficial for users of these instruments in real-world applications, such as those presented in this work.

5. Acknowledgements

JR acknowledge the financial support from the Spanish Ministry of Science, Innovation and Universities (MICIU), the State Research Agency (AEI) and the European Regional Development Fund (ERDF), EU: PID2022-136649OB-I00.

6. Declarations

The authors declare that they have no financial or non-financial competing interests related to the work submitted for publication.

Author contribution statement. Conceptualization: Barbara Giussani; Methodology: Barbara Giussani and Jordi Riu; Formal analysis and investigation: Jordi Riu and Manuel Monti; Writing - original draft preparation: Barbara Giussani; Writing - review and editing: Jordi Riu; Supervision: Barbara Giussani; Software: Jordi Riu

7. Bibliography

1. Anastas P, Eghbali N (2009) Green Chemistry: Principles and Practice. Chem Soc Rev 39:301–312. https://doi.org/10.1039/B918763B

2. Sajid M, Płotka-Wasylka J (2022) Green analytical chemistry metrics: A review. Talanta 238:123046. https://doi.org/10.1016/j.talanta.2021.123046

3. Santana-Mayor Á, Rodríguez-Ramos R, Herrera-Herrera A V., Socas-Rodríguez B, Rodríguez-Delgado MÁ (2021) Deep eutectic solvents. The new generation of green solvents in analytical chemistry. TrAC Trends in Analytical Chemistry 134:116108. https://doi.org/10.1016/j.trac.2020.116108

4. Cebi N, Bekiroglu H, Erarslan A, Rodriguez-Saona L (2023) Rapid Sensing: Hand-Held and Portable FTIR Applications for On-Site Food Quality Control from Farm to Fork. Molecules 28:1–15. https://doi.org/10.3390/molecules28093727

5. Mishra S, Singh SP, Kumar P, Khan MA, Singh S (2023) Emerging electrochemical portable methodologies on carbon-based electrocatalyst for the determination of pharmaceutical and pest control pollutants: State of the art. J Environ Chem Eng 11:109023. https://doi.org/10.1016/j.jece.2022.109023

6. Gullifa G, Barone L, Papa E, Giuffrida A, Materazzi S, Risoluti R (2023) Portable NIR spectroscopy: the route to green analytical chemistry. Front Chem 11:1–19. https://doi.org/10.3389/fchem.2023.1214825

7. Eyvazi S, Baradaran B, Mokhtarzadeh A, Guardia M de la (2021) Recent advances on development of portable biosensors for monitoring of biological contaminants in foods. Trends Food Sci Technol 114:712–721. https://doi.org/10.1016/j.tifs.2021.06.024

8. He Q, Wang B, Liang J, Liu J, Liang B, Li G, Long Y, Zhang G, Liu H (2023) Research on the construction of portable electrochemical sensors for environmental compounds quality monitoring. Mater Today Adv 17:100340. https://doi.org/10.1016/j.mtadv.2022.100340

9. Sajid M, Płotka-Wasylka J (2022) Green analytical chemistry metrics: A review. Talanta 238:123046. https://doi.org/10.1016/j.talanta.2021.123046

10. López-Lorente ÁI, Pena-Pereira F, Pedersen-Bjergaard S, Zuin VG, Ozkan SA, Psillakis E (2022) The ten principles of green sample preparation. TrAC Trends in Analytical Chemistry 148:116530. https://doi.org/10.1016/j.trac.2022.116530

11. Rinnan Å, Berg F van den, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends in Analytical Chemistry 28:1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

12. Mishra P, Biancolillo A, Roger JM, Marini F, Rutledge DN (2020) New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC - Trends in Analytical Chemistry 132:116045. https://doi.org/10.1016/j.trac.2020.116045

13. Jiao Y, Li Z, Chen X, Fei S (2020) Preprocessing methods for near-infrared spectrum calibration. J Chemom 34:e3306. https://doi.org/10.1002/cem.3306

14. Lee LC, Liong CY, Jemain AA (2017) A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum. Chemometrics and Intelligent Laboratory Systems 163:64–75. https://doi.org/10.1016/j.chemolab.2017.02.008

15. Roger JM, Biancolillo A, Marini F (2020) Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems 199:103975. https://doi.org/10.1016/j.chemolab.2020.103975

16. Schoot M, Kapper C, van Kollenburg GH, Postma GJ, van Kessel G, Buydens LMC, Jansen JJ (2020) Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample

size. Chemometrics and Intelligent Laboratory Systems 204:104105. https://doi.org/10.1016/j.chemolab.2020.104105

17. Ezenarro J, Schorn-García D, Aceña L, Mestres M, Busto O, Boqué R (2023) J-Score: A new joint parameter for PLSR model performance evaluation of spectroscopic data. Chemometrics and Intelligent Laboratory Systems 240:104883. https://doi.org/10.1016/j.chemolab.2023.104883

18. Özcan MM (2023) A review on some properties of almond: ımpact of processing, fatty acids, polyphenols, nutrients, bioactive properties, and health aspects. J Food Sci Technol 60:1493–1504. https://doi.org/10.1007/s13197-022-05398-0

19. Giussani B, Gorla G, Riu J (2024) Analytical Chemistry Strategies in the Use of Miniaturised NIR Instruments: An Overview. Crit Rev Anal Chem 54:11–43. https://doi.org/10.1080/10408347.2022.2047607

20. Ezenarro J, Riu J, Ahmed HJ, Busto O, Giussani B, Boqué R (2024) Measurement errors and implications for preprocessing in miniaturised near-infrared spectrometers: Classification of sweet and bitter almonds as a case of study. Talanta 276:126271. https://doi.org/10.1016/j.talanta.2024.126271

21. Beć KB, Grabska J, Huck CW (2021) Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. Chemistry - A European Journal 27:1514–1532. https://doi.org/10.1002/chem.202002838

22. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17. https://doi.org/10.1016/0003-2670(86)80028-9

23. Brereton RG (2000) Introduction to multivariate calibration in analytical chemistry. Analyst 125:2125–2154. https://doi.org/10.1039/b003805i

24. Bro R (2003) Multivariate calibration: What is in chemometrics for the analytical chemist? Anal Chim Acta 500:185–194. https://doi.org/10.1016/S0003-2670(03)00681-0

25. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58:109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

26. Riu J, Vega A, Boqué R, Giussani B (2022) Exploring the Analytical Complexities in Insect Powder Analysis Using Miniaturized NIR Spectroscopy. Foods 11:1–16. https://doi.org/10.3390/foods11213524

27. Gorla G, Taiana A, Boqué R, Bani P, Gachiuta O, Giussani B (2022) Unravelling error sources in miniaturized NIR spectroscopic measurements: The case study of forages. Anal Chim Acta 1211:339900. https://doi.org/10.1016/j.aca.2022.339900

28. Wentzell PD (2014) Measurement errors in multivariate chemical data. J Braz Chem Soc 25:183–196. https://doi.org/10.5935/0103-5053.20130293

29. Leger MN, Vega-Montoto L, Wentzell PD (2005) Methods for systematic investigation of measurement error covariance matrices. Chemometrics and Intelligent Laboratory Systems 77:181–205. https://doi.org/10.1016/j.chemolab.2004.09.017

30. Matinrad F, Kompany-Zareh M, Omidikia N, Dadashi M (2020) Systematic investigation of the measurement error structure in a smartphone-based spectrophotometer. Anal Chim Acta 1129:98–107. https://doi.org/10.1016/j.aca.2020.06.066

31. Westad F, Marini F (2015) Validation of chemometric models - A tutorial. Anal Chim Acta 893:14–24. https://doi.org/10.1016/j.aca.2015.06.056

32. Gorla G, Taborelli P, Giussani B (2023) A Multivariate Analysis-Driven Workflow to Tackle Uncertainties in Miniaturized NIR Data. Molecules 28:7999. https://doi.org/10.3390/molecules28247999

33. Gorla G, Taborelli P, Alamprese C, Grassi S, Giussani B (2023) On the Importance of Investigating Data Structure in Miniaturized NIR Spectroscopy Measurements of Food: The Case Study of Sugar. Foods 12:493. https://doi.org/10.3390/foods12030493

34. Wentzell PD, Wicks CC, Braga JWB, Soares LF, Pastore TCM, Coradin VTR, Davrieux F (2018) Implications of measurement error structure on the visualization of multivariate chemical data: hazards and alternatives. Can J Chem 96:738–748. https://doi.org/10.1139/cjc-2017-0730