Analytica Chimica Acta 1343 (2025) 343705

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Quantification of spectral measurement errors to guide preprocessing method selection: A case study on cannabinoid prediction across multiple NIR instruments

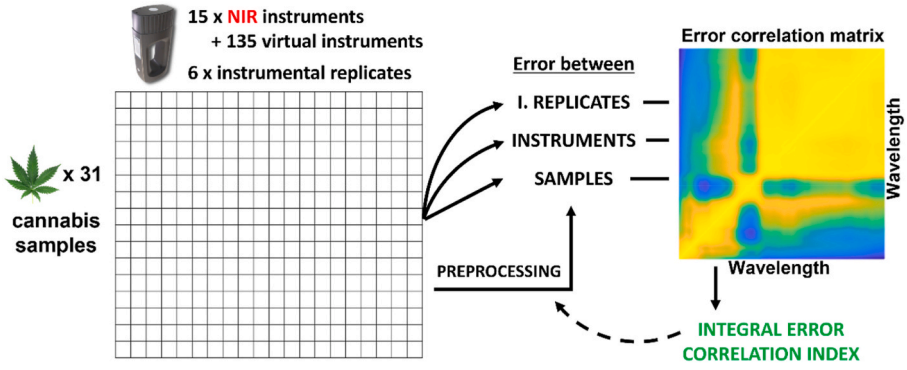
Jokin Ezenarro [a,*], Daniel Schorn-García [b], Marçal Plans [c], Olga Busto [a], Ricard Boqué [a]

[a] Universitat Rovira i Virgili, ChemoSens group, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, 43007, Tarragona, (Catalonia), Spain
[b] Stellenbosch University, Department of Viticulture and Oenology, South African Grape and Wine Research Institute, South Africa
[c] Si-ware Inc, 101 Jefferson Drive, 1st Floor, Menlo Park, CA, 94025, USA

HIGHLIGHTS

- Study addresses measurement variability across multiple NIR instruments.
- Novel IECI metric quantifies error correlation comprehensively.
- The IECI guides in preprocessing method screening and selection.
- Results enhance PLS model accuracy and reliability in cannabis analysis.
- Framework applicable for error management in diverse spectroscopy applications.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Diagonality
Preprocessing
Error covariance matrix
Error correlation
Variability sources
Heteroscedasticity

ABSTRACT

This study investigates the influence of spectral measurement errors on the accuracy and reliability of Near-Infrared (NIR) spectroscopy in predicting cannabinoid content, specifically examining the variability across multiple NIR instruments of the same model and virtual instruments. Through a detailed case study using NeoSpectra miniaturised spectrometers, we explore the sources and structures of measurement errors, their covariance and correlation patterns, the implications on preprocessing, and subsequent model performance. This study also introduces the Integral Error Correlation Index (IECI), a novel metric designed to objectively quantify measurement error correlation, as meeting the independent and identically distributed (iid) error assumption is critical for Partial Least Squares (PLS) regression models. This metric is proposed for aiding in the systematic exploration of preprocessing methods through their impact on error correlations, and their subsequent model performance. The results underscore that preprocessing methods yielding lower IECI values lead to simplified, more accurate PLS models, demonstrating the potential for improved prediction reliability. This research contributes to the optimisation of NIR spectroscopy in cannabinoid determination or other applications, offering a robust framework for managing measurement errors coming from different sources and refining multivariate predictive models in analytical methods.

* Corresponding author.
E-mail address: jokin.ezenarro@urv.cat (J. Ezenarro).

https://doi.org/10.1016/j.aca.2025.343705
Received 7 November 2024; Received in revised form 28 December 2024; Accepted 21 January 2025
Available online 23 January 2025

1. Introduction

Near-infrared (NIR) spectroscopy has rapidly emerged as an indispensable analytical tool, due to its rapid, non-destructive, and efficient measurement capabilities [1]. This technology offers promising applications in the agri-food industry as almost every organic molecule absorbs in the near-infrared region [2,3]. An example is the analysis of cannabis samples, where NIR spectroscopy enables rapid, non-destructive assessment of chemical composition. The cannabis production industry requires robust analytical methods to ensure product quality, safety, and regulatory compliance [4]. The importance of reliable and accurate analytical methods in this context cannot be overstated, given the growing legal and commercial interest in cannabis products. However, discrepancies in cannabinoid measurements between laboratories are a common issue in the cannabis industry, often due to variations in testing methods, equipment calibration, sample preparation or the instrument itself [5]. These inconsistencies can lead to unreliable reports, impacting breeders, growers, and product labelling. The lack of standardisation in testing processes, including extraction solvents and chromatographic techniques, has been particularly problematic in regions like California, where studies have revealed significant variations in tetrahydrocannabinol (THC) and cannabidiol (CBD) levels across labs [6]. These findings underscore the urgent need for standardised testing procedures to ensure accurate and consistent cannabinoid data.

In this sense, NIR spectroscopy can facilitate the control and determination of various compounds of interest in cannabis samples, englobed as cannabinoids, without the need for complex and time-consuming sample preparation [7,8] in comparison to chromatography [9]. However, despite its numerous advantages, NIR spectroscopy is not free of challenges. One of the primary concerns in its application is the presence of spectral measurement errors, which can significantly impact the accuracy and reliability of analytical models [10].

Measurement errors in spectroscopy can arise from various sources, including instrumental variability, environmental conditions, and sample preparation inconsistencies [11]. Most of the time analytical method developers are primarily concerned with having consistent and reproducible sample preparation and measurement [12]. When developing a model for commercial use, two different approaches are often used to overcome the challenges associated with inter-instrument variability and measurement inconsistencies: global model building and calibration transfer. Global models, the approach considered in this study, are designed to accommodate data variability across multiple instruments, enhancing their robustness and generalisability, which requires collecting and including all the potential variability that could be found by the model in its deployment, even including synthetic (or virtual) data generated to simulate and generalise this variability [13]. Calibration transfer techniques facilitate the application of a model developed on one instrument to others without significant loss in accuracy, thus enabling consistent performance across devices, which often require complex *ad hoc* transfer methods [14]. Incorporating such strategies can mitigate the impact of instrument-specific biases and improve the reliability of multivariate predictive models.

Independently of the strategy used for developing these commercial models, intended to be used on multiple instruments across different locations and over time, it is essential to account for the magnitude of measurement errors that may arise from variabilities between these instruments. While previous studies have documented the variability among different miniaturised NIR instruments [15–17], less attention has been given to the variability within instruments of the same model. Manufacturers often rely on users' calibration validation to ensure consistency between units [18]. However, even though the inter-instrumental variability of NeoSpectra devices has been shown [19], there is not much literature in this topic, suggesting that this issue has not been thoroughly explored. This inter-instrument variability can introduce significant discrepancies in spectral data, potentially leading to erroneous interpretations and conclusions. This variability can be influenced by slight differences in optical components, electronic noise, and manufacturing tolerances, which are even more critical in miniaturised or portable instruments [15]. Understanding and mitigating these errors is crucial for ensuring the reliability and reproducibility of NIR spectroscopy [11]. However, this aspect of instrument performance is often not disclosed by manufacturers, as their goal is to build reproducible instruments; overlooked by developers, as their goal is to offer working models; and not observed by researchers, as they usually do not have more than one instrument of the same model. Studying these errors is essential for having a deeper understanding of the fundamental limitations and potential improvements needed in the analytical methodology.

This is, the magnitude of measurement errors, their structure and sources, are key parameters to assess in analytical chemistry [16], and here comes the importance of replicates. Replicates serve as a critical tool for empirically estimating the precision of measurements, identifying potential outliers, and estimating the uncertainty associated with the different steps of the analytical process. The number and type of replicates used in an analysis can profoundly influence the robustness and reliability of the results. For instance, technical or instrumental replicates, which involve repeated measurements of the same sample, are crucial for quantifying instrument precision and detecting random errors. In this study, only this type of replicates has been considered for evaluating the instrumental measurement errors, avoiding the inclusion of variability coming from other sources.

The importance of measurement errors does not only include data acquisition, they have a big and often overlooked effect on the posterior predictions of the models built. Partial least squares (PLS) is one of the most used algorithms for regression model building; however, this algorithm assumes the errors in the data to be independent and identically distributed (known as the iid assumption). So, when the errors in the data diverge from this assumption, the performance of the model degrades, in a way that it is not obvious to the analyst that this might be the cause. The iid assumption for multivariate data can be broken in two ways: the first is whether the errors are correlated between different variables or wavelengths; and the second is heteroscedasticity, this is, unequal variance in errors across different wavelengths. In addition, measurement errors not only need to be iid, but they also need to be low. When the prediction uncertainty of the model is assessed using the error propagation approach or the error-in-variable (EIV) model, the estimation described in Equation (1) is proposed:

$$s_y^2 = SEN^{-2} s_x^2 + h\, SEN^{-2} s_x^2 + h\, s_{ycal}^2 \qquad \text{(Eq. 1)}$$

where s_y^2 is the variance of errors in prediction or prediction uncertainty, s_x^2 is the variance of the error in the signal, s_{ycal}^2 is the variance of the errors in calibration concentrations, and SEN is the sensitivity, calculated as $1/(\mathbf{b}^T\mathbf{b})^{1/2}$, with **b** being the vector of regression coefficients of the model [20]. Both the variance of errors (or uncertainty) in the signal, both for the calibration set and the sample to be predicted play a major role in the uncertainty of the prediction. And from this it can be concluded that it is of outmost importance to study the signal measurement errors in the model building process as, in the one hand, approaching iid errors will improve the fit of the model, and on the other hand, minimising the errors in the signal will significantly reduce the uncertainty in prediction.

The objective of this study is to investigate the spectral measurement errors associated with NIR spectroscopy in the analysis of cannabis, with a particular focus on the variability among instruments of the same model (NeoSpectra in this case), as it has been shown in literature that studying these errors offers valuable insights about the measurements. Recently, Gorla et al. [15] tried to reveal the error sources in one miniaturised instrument and used this information to determine different properties of forage samples. Similarly, Ezenarro et al. [21] used the measurement error study to compare different miniaturised NIR

instruments, preprocessing methods and their performance in the classification of sweet and bitter almonds. And Wentzell et al. [22] studied the error structures present in NIR spectra of wood samples for differentiating tree species. They all concluded that optimal analytical strategies can be developed by studying the error structures present in the data.

In addition, by examining the implications of these errors and their behaviour with different spectral preprocessing methodologies, the reliability and accuracy of NIR spectroscopy in this context may be enhanced. Preprocessing steps, such as baseline correction, normalisation, smoothing or derivatives, are integral to the spectral analysis process, often times they are used to improve the signal-to-noise ratio, remove unwanted systematic variance such as the one caused by light-scattering, or to enhance specific features of the data. Spectral preprocessing also has an effect on the errors of the data, potentially minimising it and making it more iid, but at the same time it is susceptible to amplifying measurement errors or removing relevant information if not appropriately addressed. This is, the preprocessing step is a critical component of spectral data modelling, and as such it must be carefully addressed, even if usually a trial-and-error approach is used for their exploration. Because of this, one of the aims of this study is to review and propose tools for an objective approach to select spectral preprocessing methods.

Ultimately, the goal of this research is to develop a more thorough understanding of the factors that influence the accuracy and reliability of NIR spectroscopic data acquisition and modelling, and to provide practical solutions for overcoming these challenges. By systematically investigating the sources of spectral measurement errors and the role of replicates, with a special look at the effects of preprocessing, this work intends to offer tools to release the full potential of NIR spectroscopy and multivariate prediction models, thereby facilitating its adoption in the cannabis industry and many other analytical settings.

2. Materials and methods

2.1. Instrumentation

Fifteen near-infrared spectrometers of the same model were used: the FT-NIR NeoSpectra Scanner spectrometer by Si-Ware Systems (Menlo Park, CA, USA). This device has the dimensions of 18.5 × 4.5 × 8 cm with a weight of approximately 730 g, and it is a handheld portable instrument. The spectra were acquired using the powder kit (Si-Ware systems), which provides boron silicate Petri dishes where samples are placed for measurement and an external blank reference tile that can accommodate the difference in height between the scanner window and the Petri dish. Boxcar Apodization function and 32,000 points for Fast Fourier Transform were used to define Fourier transform setting in the device. The instrument was adjusted to a blank reference approximately 15 min after the spectrometer was turned on, and then every 3 samples, with the 99 % reflective ceramic standard tile provided by Si-Ware. Spectra were acquired in the range of 1351–2559 nm, with 257 data points and an average spectral resolution of 4.7 nm.

Based on these measurements, virtual or synthetic instruments were simulated using the NeoSpectra Generaliser algorithm v1.0 [23]. This algorithm was built using a function set that modifies the known parameters affecting the spectra (Table 1). In summary, the algorithm inputs the spectra of the real scanners, reads the Scanner ID and creates a defined number of virtual devices based on the scanner parameters in the expected production variability, these device simulations are applied to the spectra and synthetic spectra are created. The outcomes are the spectra of the samples with small perturbations on the scanner parameters. The spectral conversion process may include multiple spectral effects such as resolution variations simulated by convolution with the instrument line shape function, wavelength errors by adding systematic or random shifts on the wavelength vector of the data, photometric errors and scaling factors, different levels of noise (additive or multiplicative correlated or uncorrelated with the signal), back-reflection and Etalon fringes, baseline shifts, self-apodization effects, absorbance scaling (light penetration depth variations), sample interface effects leading to multiple reflections of the light and environmental effects leading to temperature variations of the components.

Table 1
Characteristics of the spectral sensors extracted by the algorithm.

Characteristics	Description
SNR	Random values on y-axis with wavelength dependence
Wavelength repeatability	Random values on x-axis with wavelength dependence
Wavelength error	Shift in X-axis with zero, first or higher order dependence on wavelength
Self-apodization	Attenuation of the single wavelength line interferogram with optical path difference. The attention is a function of wavelength
Baseline shift	Shift in y-axis with zero, first or higher order dependence on wavelength
Back reflection/offset signal	Offset and scaling in y-axis with zero, first or higher order dependence on wavelength
Temp. variations	Variation in light modulation components or photodetector wavelength response versus temperature

2.2. Samples

The samples and measurements included in the dataset used for this study were extracted from a bigger dataset created by Valenveras S.L. for building cannabinoid content prediction models. Therefore, most of the provided data were not included and only the samples measured in most or all of the instruments were retained to build the dataset, as shown in Fig. 1. The set of samples consisted of 31 *Cannabis sativa* samples, dried and grinded individually before the measurements to be homogeneous in composition and particle size. The measurements were performed in random order in several analytical sessions but in continuous days, so the possible degradation of the samples was minimal. Six repeated spectral measurements were acquired each time a sample was placed on an instrument (instrumental replicates). Also, the time between the spectral measurements and posterior measurement by the reference method was minimal. The reference method to determine the cannabinoid content (THC and CBD) was High Performance Liquid Chromatography with Diode Array Detector (HPLC-DAD) as described by Aizpurua-Olaizola et al. [24] performed by Sovereign Fields S.L.

The samples used, described in Table 2, were representative for the prediction model's scope. This is, samples with high CBD content, used for medicinal purposes, samples with high THC content, used for recreative purposes, and samples with low CBD and THC content, intended for industrial use, were included in the set of samples [6]. However, the cannabinoid samples with high CBD and THC concentration are underrepresented, and low concentration samples overrepresented in the model (as shown by the median values in Table 2), thus reflecting the reality of random sampling.

2.3. Statistical data analysis

MATLAB R2022b (Mathworks Inc., Natick, MA, USA), PLS_Toolbox v9.2 (Eigenvector Inc, Manson, WA, USA) and ProSpecTool v1.0 [25] were used for data analysis. Data were organized in a structure resembling the table in Fig. 1, where each cell contained a matrix with 6 instrumental replicates × 257 wavelength variables, in order to easily restructure the data for the different studies. Error Covariance Matrices ($\mathbf{\Sigma}_{cov}$) and the Diagonality Index (DI) were calculated using in-house routines.

Different spectral preprocessing techniques were iterated, using all combinations contemplated by the ProSpecTool: gaussian smoothing, Savitzky-Golay smoothing and first and second SG derivatives (different order polynomials and window sizes), detrending, asymmetric least

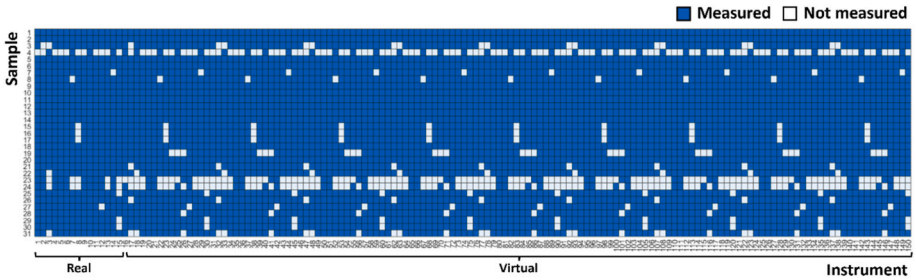


Fig. 1. Representation of the dataset containing the available measurements for each sample and instrument, where each cell represents six instrumental replicates of the NIR spectra.

Table 2
Chemical properties of the samples under study.

Compound	Range	Median	Standard deviation
Total CBD content	0.01–22.74 %	6.53 %	7.32
Total THC content	0.04–26.53 %	1.10 %	8.27

squares baseline correction, standard normal variate (SNV), autoscaling and their combinations. After spectral pre-processing, data were finally mean-centred in all calculations. The models were built after averaging the measurement replicates and validation metrics were obtained using random subsets 5-fold cross validation. The number of latent variables (LVs) for each model was automatically selected based on the minimal J-Score [26].

2.3.1. Error covariance matrix

The error covariance matrix ($\mathbf{\Sigma_{cov}}$) and its normalised form, the error correlation matrix ($\mathbf{\Sigma_{corr}}$), are a practical way to characterise and visualise multivariate measurement errors, by describing the relationships between measurement errors across the different variables, in this case, wavelengths [16,27]. Both $\mathbf{\Sigma}$ are symmetric matrices, in the case of $\mathbf{\Sigma_{cov}}$ 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 variables. For $\mathbf{\Sigma_{corr}}$ the diagonals are correlations of variables with itself (this is, equal to one) and off-diagonal elements represent the correlation with other variables. The $\mathbf{\Sigma}$ are usually graphically depicted, allowing for a visual analysis that reveals valuable insights into the magnitude and types of errors present [15,21]. This also serves as a crucial tool for identifying the underlying structure of the measured errors (such as constant or proportional errors), which can illustrate the selection of the optimal data preprocessing method. In cases where there is limited prior knowledge about the nature and structure of errors within a measurement, such as in this study involving miniaturised instruments and simulated spectra, the $\mathbf{\Sigma}$ are derived from an experimental estimation method that relies on analysing replicate measurements. This method involves estimating the (approximate) true spectrum of a sample by calculating the mean of its spectral replicates [15]. Subsequently, a residual matrix is obtained by subtracting this estimated true value from each replicate spectrum. Finally, the error covariance matrix is computed as the covariance of these residuals, as illustrated in Equation (2):

$$\mathbf{\Sigma}_{cov} = \frac{1}{(n-1)} \sum_{k=1}^{n} (\mathbf{X}_k - \overline{\mathbf{X}})^{\mathrm{T}} (\mathbf{X}_k - \overline{\mathbf{X}}) \quad \text{(Eq. 2)}$$

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 $\overline{\mathbf{X}}$ is the mean spectrum of n replicates.

However, the covariance matrix is influenced by the magnitude of the variances, which can complicate its visual interpretation, especially when few channels exhibit significantly higher variances (for instance, certain wavelengths with much lower precision compared to others). This can mask variations among other channels [15,21]. To address this issue, the covariances can be normalised to correlation values, thus, calculating the error correlation matrix by scaling the variances, eliminating the impact of varying magnitudes. The error correlation matrix, with values normalised between +1 and −1, represents the correlation coefficients derived from the covariance matrix and is calculated as shown in Equation (3).

$$\mathbf{\Sigma_{corr}} = \mathbf{\Sigma_{cov}} \cdot \Big/ \sqrt{diag(\mathbf{\Sigma_{cov}}) diag(\mathbf{\Sigma_{cov}})^{\mathrm{T}}} \quad \text{(Eq. 3)}$$

Unless the number of replicates for a sample is high enough to confidently describe the normal distribution of errors (which is not usually the case), 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 by taking the mean of all covariance matrices ($\mathbf{\Sigma_{pooled}}$). The pooling solution is generally preferred with NIR spectra because the measurement errors do not significantly change for the same types of samples [27]. In our case, covariance and correlation matrices were pooled over the different instruments, the different cannabis samples or both, depending on the objective.

2.3.2. Integral Error Correlation Index

In this work a new index is presented with the aim of simplifying and objectivising the analysis of the error correlation matrices. This index, the Integral Error Correlation Index (IECI), is a comprehensive metric that seeks to quantify the correlation between errors in the dataset with a single value.

For the calculation of the IECI, firstly, the Error Correlation Distribution (ECD) is calculated. This is a vector containing the sum of values of the off-diagonals in the correlation matrix, as a function of the distance from the diagonal of the matrix (illustrated in Fig. S1). It can be mathematically described as shown in Equation (4):

$$\boldsymbol{ECD}(i) = \sum_{j=1}^{i} |\mathbf{\Sigma_{corr}}(j, i-j+1)| \quad \text{(Eq. 4)}$$

Then, based on the ECD, the IECI is calculated as a ratio of areas under the curve between the area of the ECD and the area of what would be the ECD of a fully correlated matrix. This is equivalent to the average value of the elements of the $\mathbf{\Sigma}_{\mathrm{corr}}$ after removing the diagonal, which can be mathematically described as shown in Equation (5):

$$IECI = \frac{\sum_{i=1}^{N-1} \boldsymbol{ECD}(i)}{\sum_{i=1}^{N-1} i} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{\Sigma_{corr}}(i,j) - \sum_{i=1}^{N} \mathbf{\Sigma_{corr}}(i,i)}{N^2 - N} \quad \text{(Eq. 5)}$$

This is an index that ranges between 0 and 1; the higher the IECI, the closer the correlation distribution to a matrix with fully correlated errors (non-iid); the lower the IECI, the closer the correlation distribution to a matrix with fully random errors (iid).

2.3.3. Partial Least Squares regression (PLSR)

PLS regression is a supervised multivariate method used to model the relationship between a set of predictor variables (X) and response

variables (Y). PLS regression reduces the dimensionality of the data by projecting both X and Y onto a new set of latent variables, called factors, which capture the most relevant variance in X for predicting Y. In this study, PLS regression was employed to model the relationship between the NIR spectra of cannabis samples (X) and the cannabinoid content (Y), and the number of latent variables was chosen taking into consideration the minimum value for the J-score, as explained in Ezenarro et al. [26].

3. Results and discussion

3.1. Study of error sources

When a spectral dataset is obtained, different properties can be explored considering different types of replicates. This is, if only instrumental replicates of the same sample are considered, the only error shown is the instrumental. However, by comparing same samples in different instruments or same instruments in different samples the population of instruments or population of samples can be explored, respectively. And this can be done using error covariance matrices.

The first thing to deduce from Fig. 2 is that error covariance matrices are complex but can offer considerable information; therefore, they can be used as a basic data exploratory tool. It can also be concluded that, as expected, if many more virtual than real instruments are included in the dataset, the characteristics of the virtual instruments will overcome the real data and their properties, even in the error (third column in Fig. 2). Therefore, it is paramount to make sure that the real and simulated measurements have similar properties. By looking at the errors between instrumental replicates (first row of Fig. 2), it can be seen that they have quite similar but not equal structures. All of them present a high correlation peak in the centre of the spectra, which is related to a water band. This reflects the issue that as the sample is measured, it is heated by the NIR light source, causing systematic variance on this band. On top of that, the virtual instrument replicates have high variance at the beginning of the spectra (lower left corner), which may come from some sort of extrapolation, also called "tail effect", and in general, causing higher variance and covariance values than the real measurements in this part. In practice, this would mean that the optimal preprocessing method (the preprocessing method that equalises instrumental replicates) may not be the same for real or virtual instruments; and considering all together, as virtual instruments have a higher weight due to their number, the method that minimises errors caused by this "tail effect" may prevail, even if in real instruments it is not needed.

Regarding the errors between instruments considering the same samples (second row of Fig. 2), both real and virtual instruments have the same structure and scale in the covariance matrix. This was expected, as the generalisation algorithm used to simulate the new instruments emulates the population of real instruments. In this sense, the generaliser works, and the preprocessing would affect real and virtual instruments in the same way. However, by looking at the error between samples (third row of Fig. 2), it can be seen that real and virtual instruments covariance matrices differ in both structure and scale. It must be noted that while the previously discussed matrices only contain variability coming from systematic and/or random sources, the variability described in these matrices is also related to chemical information, this is, the differences in sample composition. It can be seen that even if the underlying structure may be similar, the virtual instruments add a significant amount of variability at the end of the spectra (top right corner), which cannot be related to the chemical properties of the samples. Therefore, it can be concluded that this generalisation algorithm (for this experimental design and using this version of the generaliser) does not preserve the properties of the measurements in the variability augmentation process and that the real and virtual instruments may offer non-equivalent measurements. And as stated above, using more virtual instruments than real ones overcomes their characteristics, so, for the proper characterisation of the measurement errors and their implications in preprocessing and subsequent modelling, only real instrument measurements will be considered from now on. This proves the use of error covariance matrices as a basic data exploration tool.

Coming to the real instruments, even if the manufacturers state similar quality and characteristics for all the instruments of the same

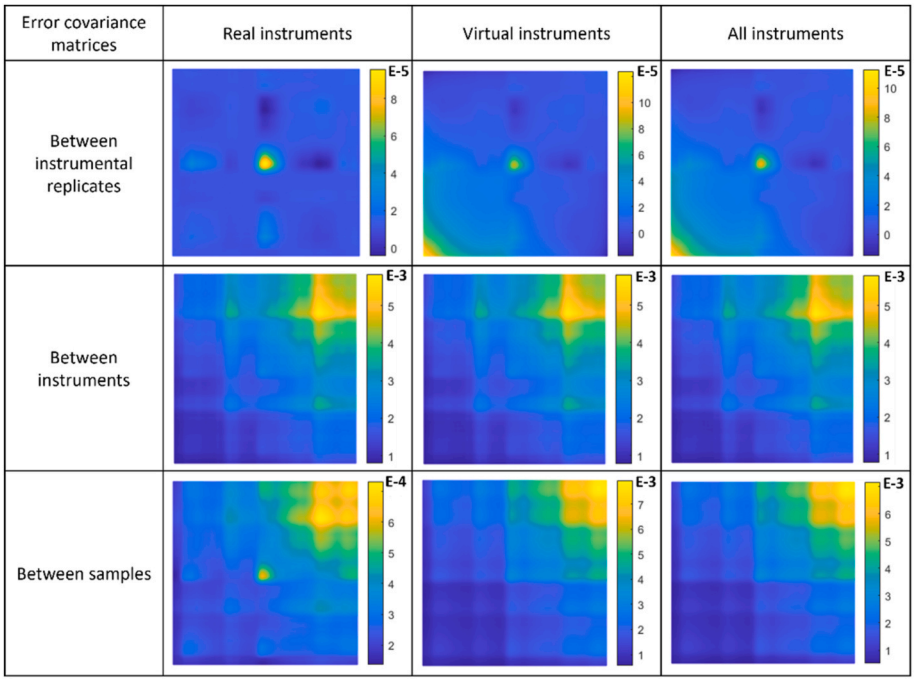


Fig. 2. Error covariance matrices for the real, virtual, and real and virtual instruments considering different types of replicates: instrumental, between instruments, and between samples; pooling different matrices for each purpose. Both axes represent wavelengths (1351–2559 nm), increasing from left to right and from bottom to top.

model, variations in producing and assembling the components may lead to variations in the measurement and performance of the instruments [19]. As the 15 instruments used in this experiment were the most extreme production cases found by the producer, to explore their differences, covariance and correlation matrices for the instrumental replicate errors were calculated for each of the 15 individual instruments. Covariance matrices (Fig. S2) show that the scale of the errors is similar for each instrument, however, Instruments 1 and 13 show unexpected variance at the errors of the final part of the spectra. Looking at the correlation matrices (Fig. 3) it can be seen that there are different structures in the correlation matrices for the different real instruments. For instance, instruments, 2, 3, 6, 11, 12 and 15 have similar structures, even though there are differences in scale. And the other instruments have different or additional error structures.

As a final point for the discussion of the measurement errors, described in both Figs. 2 and 3, it must be noted that even though they primarily highlight inter-instrument differences, other factors such as experimental procedures, equipment stability, environmental conditions, and operator handling, can also contribute to the observed bias. Differences in sample preparation, environmental fluctuations, and equipment drift may further affect data consistency. Here instrument variability is discussed as these additional factors were minimised through standardised procedures and stable measurement environment, which prove crucial for improving calibration model robustness and reliability over time.

3.2. Implications of preprocessing on errors and modelling

The preprocessing of spectral data is a fundamental step in the development of analytical methods, where measurement conditions may significantly affect the registered spectra. Selecting an appropriate preprocessing method can have varying impacts on the measurement errors inherent in the data, complicating the subsequent data analysis. Errors can propagate through the analysis pipeline leading to biased models and unreliable predictions, so effective preprocessing should reduce these errors and enhance chemical information, improving model performance.

3.2.1. Development of a metric

Preprocessing methods, as data transformation methods, have an impact on the measurement errors, and therefore, in error covariance and correlation matrices. Some researchers have relied on visual inspection of these matrices based on trial-and-error approaches to evaluate the impact of different preprocessing techniques [15,21], which is a time-consuming and often subjective approach. While these methods can provide useful insights, they are inherently limited by their reliance on human interpretation, which can introduce bias and reduce reproducibility. Moreover, this approach is impractical when many preprocessing methods are being compared.

To address these challenges, there is a need for the development of an objective, quantitative metric that can systematically evaluate the implications of preprocessing or any other change on the measurement errors, and therefore, the effectiveness of these preprocessing methods and potentially the identification of the ideal one.

The key aspect that this metric should capture is the degree of correlation of the errors in the data. Because, as explained, the correlation of errors will deteriorate the performance of regression models such as PLS. This is, the metric should therefore quantify the degree of diagonalisation achieved in the error covariance matrix, as a diagonal matrix (where off-diagonal elements are minimised) indicates that errors are uncorrelated and the data is less prone to model bias. However, as per its nature, NIR signals have high multicollinearity, this is, there is a high correlation between neighbour wavelengths. This should be taken into consideration if the error correlations are of close-by channels, meaning that they belong to the same peak and could be a systematic shift, or if they are distanced, meaning that they should be uncorrelated. To explore this, the error correlations shown in the $\mathbf{\Sigma}_{corr}$ can be added together as a function of their distance to the diagonal of the matrix, mathematically described in Equation (4) and defined as Error Correlation Distribution (ECD), shown in Fig. 4. This is, correlations are summed based on the distance between the two correlated wavelengths; close neighbours are expected to have higher correlations than distanced neighbours.

On the one hand, if the errors were fully correlated, the ECD curve would be totally diagonal (shown in Fig. 4 as a black dotted line), as the maximum value that can be achieved for each off-diagonal is the number

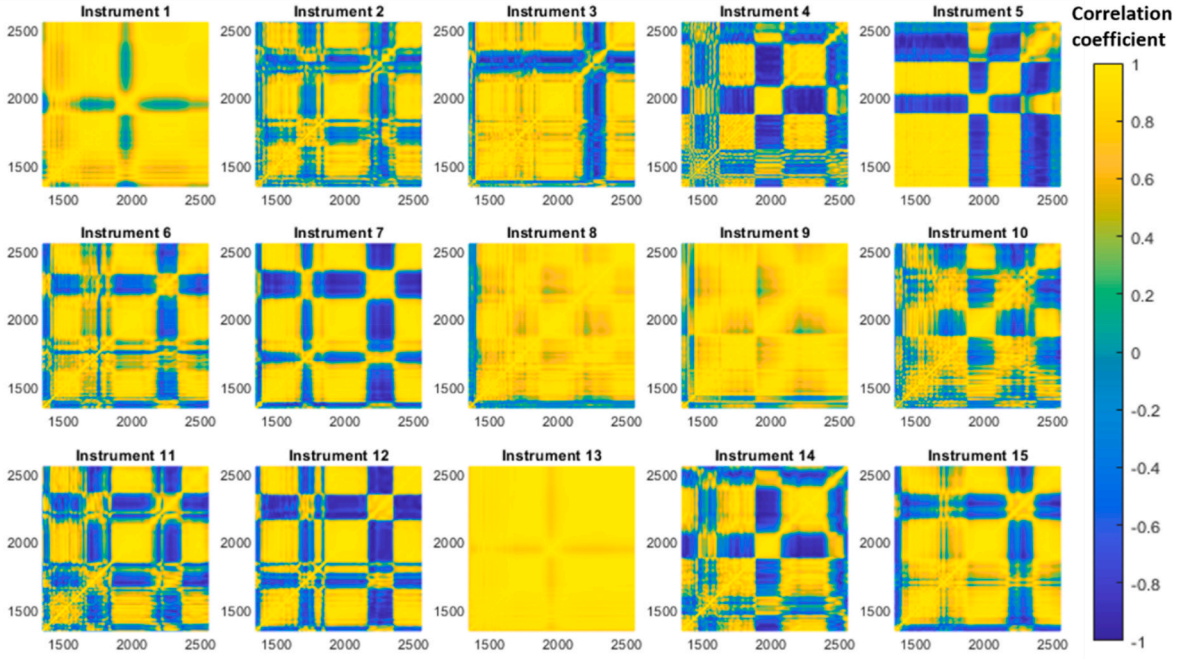


Fig. 3. Correlation matrices for the errors between instrumental replicates for each of the 15 real instruments, calculated with the same 31 samples. Axes correspond to wavelengths (nm).

of elements in that off-diagonal. On the other hand, if the errors were totally independent, the ECD curve would lay as a constant line on zero, which is virtually impossible due to the nature of NIR signals. As it can be seen in Fig. 4, the datasets after different preprocessing methods can have very different ECD profiles, going from a low almost asymptotic line to a high almost diagonal line. And even if observing at the ECD profiles facilitates the exploration of the effect of different preprocessing methods on the measurement errors, this still relies on a subjective evaluation. To convert these ECD profiles into a quantitative metric, the area under these curves can be used. If the area (sum of values) of the ECD curves is compared to the area under the diagonal, a metric can be proposed to quantify how close a certain ECD corresponding to a certain preprocessing method is to the totally correlated errors situation. This metric, which is equivalent to the average value of the $\mathbf{\Sigma}_{corr}$ after removing the diagonal, is mathematically described in Equation (5), defined as the Integral Error Correlation Index (IECI). Each ECD curve in Fig. 4 is coloured based on their IECI value: the lower the IECI value the closer to the independent error situation, and therefore, a potentially better preprocessing method.

Moreover, while it is essential to reduce errors and correlations, it is equally important that the preprocessing method preserves the informative signal within the data. Excessive smoothing or aggressive baseline correction, for example, can diminish the spectral features that are critical for accurate prediction models. This balance between noise reduction and signal retention is vital for maintaining the integrity of the data and ensuring that the models built on these data are both accurate and reliable.

3.2.2. *The IECI as a model complexity and performance indicator*

The correlation of measurement errors and their structure directly affects regression algorithms like PLS (that assumes the errors to be iid), and can significantly impact the accuracy, robustness, and predictive power, this is, the performance of the resulting models [20]. Therefore, the IECI, as a quantitative metric of errors correlation, could be an indicator of the performance of the subsequent model. This index could serve as a preliminary tool for evaluating the effectiveness of different preprocessing methods in terms of their ability to reduce error correlations and, by extension, improve model performance. For instance, preprocessing methods that produce data with low IECI values are more likely to result in simpler, more robust PLSR models with fewer latent variables, as the covariance structures are easier to explain. These models are not only easier to interpret but also more likely to generalise well to new data, which is a key consideration in the development of reliable analytical methods.

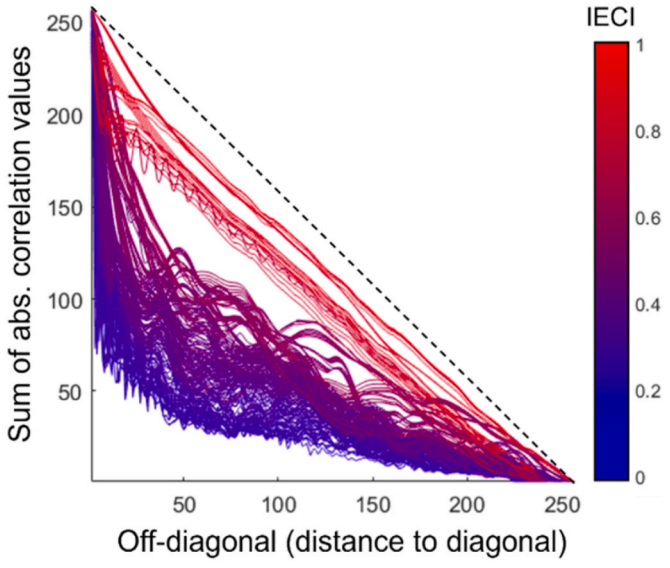


Fig. 4. The ECD curves of the datasets preprocessed with different methods, coloured by their IECI. The ECD profile of a fully correlated $\mathbf{\Sigma}_{corr}$ is represented in a black dotted line.

To explore the relation between the IECI and the model performance, the $RMSE_{CV}$ of each PLS model built for each preprocessing method was plotted against their IECI, for different number of LVs (coloured based on the optimal number of LV selected using the J-score [26]) for total CBD prediction, shown in Fig. 5 (and for THC in Fig. S3).

As these figures show, the fulfilment of the iid assumption (lower IECI) is related to the simplicity of the regression models, even if it is not a perfect correlation. This is, the higher the IECI, the more components needed to model the data and to obtain an ideal or equivalent performance, as already described by Allegrini and Olivieri [20]. This postulates the IECI as an exploration tool for raw and preprocessed data before modelling and before looking at its correlation with the dependant variable. This is precisely one of the strong and weak points of the IECI at the same time: not needing to consider the dependant variable makes this an ideal tool for studying properties of the measured data independently of the model; however, not considering the dependant variable means that the data properties may have different implications on the modelling of different parameters.

To better study this relationship between the IECI, model performance and used preprocessing methods, these are represented in Fig. 6 for the first LV (and the same for THC prediction models in Fig. S4). Before discussing this relationship, it must be noted that this dataset, as determined by the methods described in Ezenarro et al. [25], is noisy and both additive and multiplicative scatter effects are present. This means that much of the variance in the raw data comes from sources not related to the chemical properties of the samples. Usually, several preprocessing methods are needed to minimise this trivial variance and reveal the chemical information. The first conclusion is that there are two groupings, one on the left that has a high IECI vs RMSE correlation, and another one in the right, not showing this pattern. For the first group it can be concluded that derivatives play a major role in this dataset, as they remove noise and increase variations describing peak features, the error correlation (IECI) decreases, also enhancing the chemical information, thus, forming this correlation. Looking at Fig. 6b, instead, the reason of the second grouping is revealed: these are the models that include the AsLS baseline correction with a second derivative, that instead of enhancing chemical information, may also enhance correlated noise. However, the rest of sample-wise normalisations do not differ from the tendency described before, each of them forms a line correlating the IECI and the RMSE and there is not much difference between them. This is expected, as SNV and detrending, for instance, only aim at removing the scattering effects of the data.

However, the conclusion is not necessarily that the IECI predicts the performance of the model, even if for this dataset they are correlated. This could mean that this dataset, having so much trivial variance (not caused by the chemical composition of the samples), benefitted from heavy preprocessing, and this at the same time reduces the IECI, and reveals the chemical information, improving performance (RMSE). But this may only be true for the preprocessing methods that really enhance chemical information and not noise or unrelated variance, each case should be thoroughly explored.

This also suggests that the IECI is not just a data exploration tool, the IECI can also be integrated into the model selection process. During the initial stages of model development, different preprocessing methods can be applied to the spectral data, and the resulting IECI values can be calculated. Preprocessing methods that minimise the IECI should be prioritized for further model development, as they are likely to reduce the propagation of measurement errors and improve the overall model performance. This approach allows researchers to systematically and objectively select preprocessing methods that optimise the balance between error reduction and signal preservation, leading to more accurate and reliable predictive models.

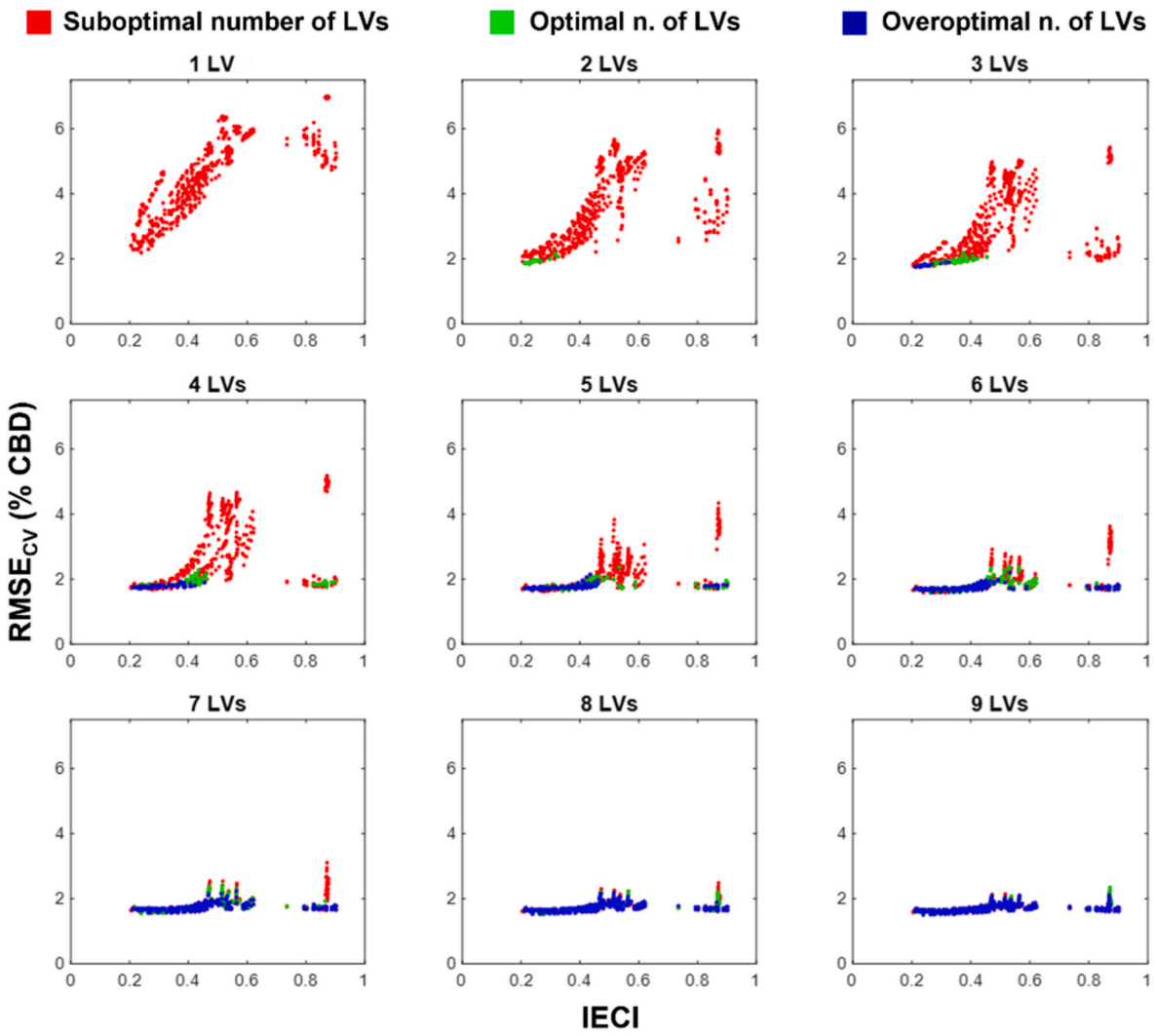


Fig. 5. The performance ($RMSE_{CV}$) of total CBD content prediction PLS models using different preprocessing methods, against their IECI, with different number of LVs. In each subplot, colours represent if that number of LVs is optimal or not for the PLS model built with each preprocessed dataset, based on the J-Score [26]. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

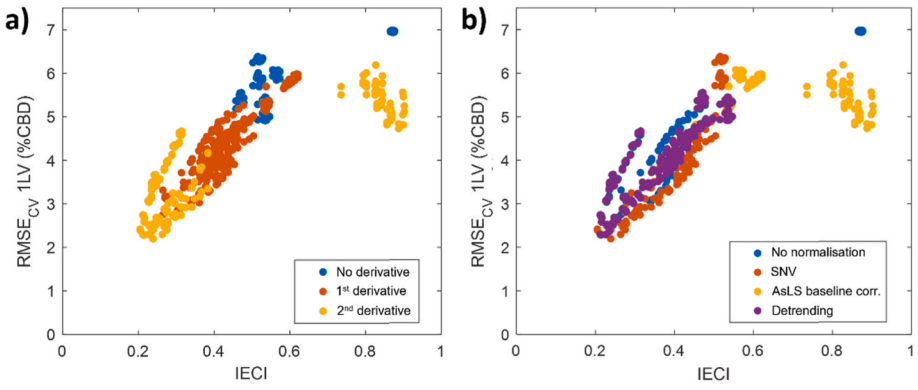


Fig. 6. The performance ($RMSE_{CV}$) of total CBD content prediction PLS models using different preprocessing methods, against their IECI, with one LV. Each point represents a specific preprocessing strategy, coloured by a) the order of the derivative or b) sample-wise normalisation algorithm.

In addition to guiding the selection of preprocessing methods, the IECI could potentially be used to monitor model performance during the calibration process. As models are iteratively refined and validated, changes in the IECI can indicate whether the preprocessing and modelling steps are effectively managing the error structures in the data. For example, if the IECI increases during model development, it may signal that the current preprocessing method is not adequately addressing the error correlations, prompting a re-evaluation of the preprocessing strategy.

Table 3
Properties of the models chosen as optimal for the prediction of CBD and THC content in cannabis. [a]Ratio of Performance to Deviation (values above 3 are considered excellent predictive performance). [b]Range Error Ratio (values above 10 are considered good predictive performance).

Model	Preprocessing	LVs	IECI	$RMSE_{CV}$	J-Score	R^2_{CV}	RPD[a]	RER[b]
CBD	Smoothing (SG 2o 13p) 2nd Derivative Detrending (1o) MC	6	0.21	1.70 %	0.31	0.95	4.31	13.37
THC	Smoothing (SG 2o 21p) 1st Derivative Detrending (2o) MC	3	0.21	2.32 %	0.27	0.93	3.56	11.42

3.2.3. *Characteristic of the selected models*

Lastly, as data exploration, modelling and model selection steps have been described, the optimal models for total CBD and THC content prediction are shown in Table 3. These models were selected based on the metrics shown in Table 3 as well as previous expertise about the data. As expected, the use of several preprocessing methods produces the simplest (in terms of dimensionality) yet well performing models. Both models show a low IECI (among the lowest 5 %), low J-Score (within 0.05 points of the top J-Score models, which used 10 LVs), low RMSE (also not differing from the top performing but more complex models), and also a high determination coefficient. Looking both at the Ratio of Performance to Deviation (RPD) and Range Error Ratio (RER), both models are considered to have good predictive performance [28].

The results of this study demonstrate comparable predictive performance and superior error metrics for both CBD and THC quantification compared to those reported in recent literature employing NIR spectroscopy for cannabinoids. Notably, the prediction model for total CBD content achieved an RPD of 4.31, indicating exceptional predictive accuracy and error minimization. This significantly surpasses the RPD values reported by Jarén et al. (2022) [29], who achieved an RPD greater than 2 for CBD quantification using NIR spectroscopy. Similarly, our THC model (RPD of 3.56) aligns closely with the findings of Su et al. (2022) [7], who reported an RPD of around 3 for total THC quantification in ground hemp samples.

The determination coefficients (R^2_{CV}) of 0.95 for CBD and 0.93 for THC in the presented models underscore a high level of calibration accuracy, consistent with the values reported by Yao et al. (2022) [30], who achieved R^2 values ranging from 0.91 to 0.95. Additionally, the low $RMSE_{CV}$ values achieved in this study (Table 3), compared to those presented in literature [7,29,30], reflect exceptional predictive capabilities. Furthermore, the minimised IECI indicates consistent model performance across varying instruments and sample conditions, significantly reducing the effects of measurement errors in practical applications.

4. Conclusions

This study has underscored the critical importance of understanding and addressing spectral measurement errors in the application of NIR spectroscopy, particularly within the cannabis industry. These errors, which can be attributed to many factors including instrumental variability, environmental factors, and sample preparation inconsistencies, have a significant impact on the accuracy and reliability of the resulting analytical models. Through the exploration of error covariance and correlation matrices, this work was able to provide deeper insights into the nature and magnitude of measurement errors, highlighting the need for a more systematic approach to error analysis. This allowed for an objective comparison between characteristics of different infrared instruments of the same model, as well as an evaluation of a generalisation algorithm to introduce variance that could be found by the model in its deployment.

One of the most significant contributions to this approach is the introduction of the Integral Error Correlation Index (IECI), which provided an objective metric to quantify the correlation of measurement errors in the data, instead of just matrix visualisation as the literature suggests. This index allowed for a clear comparison of the impact of different preprocessing methods on the error structure, offering a valuable tool for potentially elucidating best performance models. The study demonstrated that preprocessing methods that produced data with lower IECI values resulted in simpler, more robust PLS regression models, as a consequence of the iid assumption fulfilment. This insight is crucial, as it delves into the relationship between the efficacy of preprocessing and the performance of the resulting model, offering a more systematic approach to model development.

In summary, this study provides a comprehensive framework for addressing spectral measurement errors in the acquisition of the spectra and selecting optimal preprocessing methods in NIR spectroscopy. The use of automated tools like the ProSpecTool, combined with the introduction of the IECI, offers a powerful new approach to improving model performance and reducing the uncertainty associated with spectral analysis. These advancements are set to facilitate the adoption of NIR spectroscopy in various industries, including the fast-growing cannabis industry. They will contribute to the development of more accurate and reliable analytical methods that can meet the demands of both regulatory authorities and consumers, as shown by the resulting cannabinoid prediction models.

CRediT authorship contribution statement

Jokin Ezenarro: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Schorn-García:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Marçal Plans:** Writing – review & editing, Validation, Resources. **Olga Busto:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Ricard Boqué:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Funding

Grant URV Martí i Franqués – Banco Santander (2021PMF-BS-12). Chemometrics and Sensorics for Analytical Solutions (CHEMOSENS, ref.2021 SGR 00705, Department de Recerca i Universitats, Generalitat de Catalunya).

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.

Acknowledgments

The authors want to express the gratitude to Valenveras S.L. for providing the dataset used in this article.

During the preparation of this work, the authors used ChatGPT 4o in order to improve the understandability and conciseness of the text. After using this tool, the authors reviewed and edited the content as needed and took 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.aca.2025.343705.

Data availability

The authors do not have permission to share data.

References

[1] B. Giussani, G. Gorla, J. Riu, Analytical chemistry strategies in the use of miniaturised NIR instruments: an overview, Crit. Rev. Anal. Chem. 54 (2024) 11–43, https://doi.org/10.1080/10408347.2022.2047607.

[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] 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.

[4] R.J. Pusiak, C. Cox, C.S. Harris, Growing pains: an overview of cannabis quality control and quality assurance in Canada, Int. J. Drug Pol. 93 (2021) 103111, https://doi.org/10.1016/j.drugpo.2021.103111.

[5] B.C. Smith, Inter-lab variation in the cannabis industry, Part I: problem and causes, Cannabis Sci.Technol. 2 (2019) 12–17.

[6] M.J. Zoorob, The frequency distribution of reported THC concentrations of legal cannabis flower products increases discontinuously around the 20% THC threshold in Nevada and Washington state, J Cannabis Res 3 (2021) 6, https://doi.org/10.1186/s42238-021-00064-2.

[7] K. Su, E. Maghirang, J.W. Tan, J.Y. Yoon, P. Armstrong, P. Kachroo, D. Hildebrand, NIR spectroscopy for rapid measurement of moisture and cannabinoid contents of industrial hemp (*Cannabis sativa*), Ind. Crops Prod. 184 (2022) 115007, https://doi.org/10.1016/j.indcrop.2022.115007.

[8] 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.

[9] L. Nahar, A. Onder, S.D. Sarker, A review on the recent advances in HPLC, UHPLC and UPLC analyses of naturally occurring cannabinoids (2010–2019), Phytochem. Anal. 31 (2020) 413–457, https://doi.org/10.1002/pca.2906.

[10] 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.

[11] 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.

[12] E. Bouveresse, D.L. Massart, Standardisation of near-infrared spectrometric instruments: a review, Vib. Spectrosc. 11 (1996) 3–15, https://doi.org/10.1016/0924-2031(95)00055-0.

[13] M. Wohlers, A. McGlone, E. Frank, G. Holmes, Augmenting NIR Spectra in deep regression to improve calibration, Chemometr. Intell. Lab. Syst. 240 (2023) 104924, https://doi.org/10.1016/j.chemolab.2023.104924.

[14] T. Fearn, Standardisation and calibration transfer for near infrared instruments: a review, J. Near Infrared Spectrosc. 9 (2001) 229–244, https://doi.org/10.1255/jnirs.309.

[15] 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.

[16] 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.

[17] 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.

[18] PerkinElmer, Validation – adjustment of NIR calibrations, Science With Purpose (2020).

[19] S.M. Mitu, C. Smith, J. Sanderman, R.R. Ferguson, K. Shepherd, Y. Ge, Evaluating consistency across multiple NeoSpectra (compact Fourier transform near-infrared) spectrometers for estimating common soil properties, Soil Sci. Soc. Am. J. 88 (2024) 1324–1339, https://doi.org/10.1002/saj2.20678.

[20] 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.

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

[22] 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.

[23] Y.M. Sabry, B. Mortada, S. Abozyd, M. Medhat, M. Said, B. Saadany, Y. Helmy, A. Badr, M. Plans Pujolras, Generalized Artificial Intelligence Modeler for Ultra-wide-scale Deployment of Spectral Devices, 2023. US20230304860A1.

[24] O. Aizpurua-Olaizola, U. Soydaner, E. Öztürk, D. Schibano, Y. Simsir, P. Navarro, N. Etxebarria, A. Usobiaga, Evolution of the cannabinoid and terpene content during the growth of *Cannabis sativa* plants from different chemotypes, J. Nat. Prod. 79 (2016) 324–331, https://doi.org/10.1021/acs.jnatprod.5b00949.

[25] J. Ezenarro, D. Schorn-García, O. Busto, R. Boqué, ProSpecTool: a MATLAB toolbox for spectral preprocessing selection, Chemometr. Intell. Lab. Syst. 247 (2024) 105096, https://doi.org/10.1016/j.chemolab.2024.105096.

[26] J. Ezenarro, D. Schorn-García, L. Aceña, M. Mestres, O. Busto, R. Boqué, J-Score: a new joint parameter for PLSR model performance evaluation of spectroscopic data, Chemometr. Intell. Lab. Syst. 240 (2023) 104883, https://doi.org/10.1016/j.chemolab.2023.104883.

[27] 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.

[28] T. Fearn, Assessing calibrations: SEP, RPD, RER and R2, NIR News 13 (2002) 12–13, https://doi.org/10.1255/nirn.689.

[29] C. Jarén, P.C. Zambrana, C. Pérez-Roncal, A. López-Maestresalas, A. Ábrego, S. Arazuri, Potential of NIRS technology for the determination of cannabinoid content in industrial hemp (*Cannabis sativa* L.), Agronomy 12 (2022) 938, https://doi.org/10.3390/agronomy12040938.

[30] S. Yao, C. Ball, G. Miyagusuku-Cruzado, M.M. Giusti, D.P. Aykas, L.E. Rodriguez-Saona, A novel handheld FT-NIR spectroscopic approach for real-time screening of major cannabinoids content in hemp, Talanta 247 (2022) 123559, https://doi.org/10.1016/j.talanta.2022.123559.